# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 12-04-2012 | Final Report | 12-Aug-2008 - 11-Aug-2011 |

**4. TITLE AND SUBTITLE**
Final project report for "Neuroscience-enabled complex visual scene understanding"

**5a. CONTRACT NUMBER**
W911NF-08-1-0360

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
611102

**6. AUTHORS**
Laurent Itti, Nader Noori, Lior Elazary

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**
University of Southern California
Contracts and Grants
University of Southern California
Los Angeles, CA 90089 -0701

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

**10. SPONSOR/MONITOR'S ACRONYM(S)**
ARO

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
54189-NS.7

**12. DISTRIBUTION AVAILIBILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**
We have developed a new Bayesian framework for visual perception. The framework makes use of bottom-up computation heuristics (including salience maps) and top-down knowledge (where high-level hypotheses guide low- level visual processing). As this yields complex computations and a large search space of hypotheses for interpretation of the visual data, we developed a number of new techniques to make the system computationally tractable. In particular, we use probabilistic techniques reminiscent of recent approaches to probabilistic robotics

**15. SUBJECT TERMS**
neuroscience, vision, Bayesian, attention, cognition

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Laurent Itti |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER: 213-740-3527 |

**Report Title**

Final project report for "Neuroscience-enabled complex visual scene understanding"


**ABSTRACT**

We have developed a new Bayesian framework for visual perception. The framework makes use of bottom-up computation heuristics (including salience maps) and top-down knowledge (where high-level hypotheses guide low- level visual processing). As this yields complex computations and a large search space of hypotheses for interpretation of the visual data, we developed a number of new techniques to make the system computationally tractable. In particular, we use probabilistic techniques reminiscent of recent approaches to probabilistic robotics (including MCMC, DDMCMC, and particle filters). In addition, we have completed experiments to elucidate the relationship between cognition and visual processing. This work provides important guidelines for further development of our computational vision frameworks. The key question addressed here is how humans may re-use brain regions evolutionarily associated with some form of processing (e.g., vision) to serve other forms of processing (e.g., algebra, mental memorization and sorting of strings of numbers) which are too recent on an evolutionary time scale to have dedicated brain areas. This report describes both project and many applications to robotics, machine vision, and others.


---

## Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

### (a) Papers published in peer-reviewed journals (N/A for none)

| Received | Paper |
|---|---|
| 2012/04/12 15 4 | Zhicheng Li, Shiyin Qin, Laurent Itti. Visual attention guided bit allocation in video compression, Image and Vision Computing, (1 2011): 0. doi: 10.1016/j.imavis.2010.07.001 |
| 2012/04/12 15 2 | Farhan Baluch, Laurent Itti, Michael H. Herzog. Training Top-Down Attention Improves Performance on a Triple-Conjunction Search Task, PLoS ONE, (2 2010): 0. doi: 10.1371/journal.pone.0009127 |
| 2012/04/12 14 1 | Farhan Baluch, Laurent Itti. Mechanisms of top-down attention, Trends in Neurosciences, (04 2011): 0. doi: 10.1016/j.tins.2011.02.003 |

**TOTAL: 3**

**Number of Papers published in peer-reviewed journals:**

---

### (b) Papers published in non-peer-reviewed journals (N/A for none)

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Papers published in non peer-reviewed journals:**

---

### (c) Presentations

**Number of Presentations:** 0.00

---

### Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):**

## Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

**TOTAL:**

**Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):**

## (d) Manuscripts

Received Paper

**TOTAL:**

**Number of Manuscripts:**

## Books

Received Paper

**TOTAL:**

## Patents Submitted

## Patents Awarded

## Awards

## Graduate Students

| NAME | PERCENT_SUPPORTED | Discipline |
|---|---|---|
| Lior Elazary | 0.50 | |
| Nader Noori | 0.50 | |
| **FTE Equivalent:** | **1.00** | |
| **Total Number:** | **2** | |

## Names of Post Doctorates

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Names of Faculty Supported

| NAME | PERCENT_SUPPORTED | National Academy Member |
|---|---|---|
| Laurent Itti | 0.08 | No |
| **FTE Equivalent:** | **0.08** | |
| **Total Number:** | **1** | |

## Names of Under Graduate students supported

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

| | |
|---|---|
| The number of undergraduates funded by this agreement who graduated during this period: ..... | 0.00 |
| The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields: ..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields: ..... | 0.00 |
| Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale): ..... | 0.00 |
| Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering: ..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense ..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: ..... | 0.00 |

## Names of Personnel receiving masters degrees

| NAME |
|---|
| **Total Number:** |

## Names of personnel receiving PHDs

| NAME |
|---|
| **Total Number:** |

## Names of other research staff

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Sub Contractors (DD882)

## Inventions (DD882)

## Scientific Progress

see attachment

## Technology Transfer

# Final project report

**Proposal Number: 54189-NS, Agreement Number: W911NF-08-1-0360**

## "Neuroscience-enabled complex visual scene understanding"

**P.I.: Laurent Itti, University of Southern California**

**Problem statement and summary of most important results**: In this project, we have developed a new Bayesian framework for visual perception. The framework makes use of bottom-up computation heuristics (including salience maps) and top-down knowledge (where high-level hypotheses guide low-level visual processing). As this yields complex computations and a large search space of hypotheses for interpretation of the visual data, we developed a number of new techniques to make the system computationally tractable. In particular, we use probabilistic techniques reminiscent of recent approaches to probabilistic robotics (including MCMC, DDMCMC, and particle filters). This project is described in details in the following pages.

In addition, we have completed experiments to elucidate the relationship between cognition and visual processing. This work provides important guidelines for further development of our computational vision frameworks. The key question addressed here is how humans may re-use brain regions evolutionarily associated with some form of processing (e.g., vision) to serve other forms of processing (e.g., algebra, mental memorization and sorting of strings of numbers) which are too recent on an evolutionary time scale to have dedicated brain areas. The following pages also present the results from this project in details.

Applications of our new Bayesian work on attention have been rich and diverse. First, we have applied our model to search and classification of targets (Li & Itti, IEEE Trans Image Proc 2011), then to the better prediction of human eye movements when they watch complex dynamic scenes by using top-down templates on attention (Li et al., Image & Vision Computing 2011), combined bottom-up and top-down prediction of gaze and actions in interactive tasks such as driving (Borji & Itti CVPR 2012, Borji et al., CVPR 2012), top-down-guided visual search (Elazary & Itti, Vision Res 2010; Baluch & Itti PlosOne 2012), robotics (Siagian et al J Field Robotics 2011), and attention-based visual prostheses (Parikh et al J Neural Engineering 2010).

Finally, our work on this project has allowed us to write two review papers, one of which in a high-impact journal (Baluch & Itti, Trends in Neuroscience 2011, impact factor 13.3; Borji & Itti IEEE PAMI 2012).

Table of contents:


1. Publications 2
2. New Bayesian framework for perception 5
3. Neural basis of mental cognition 30
4. Appendix: published/accepted paper reprints 40

**Publications resulting from the award:**

*** From the two students directly supported by the project (L. Elazary and N. Noori):**

L. Elazary, L. Itti, A Bayesian model for efficient visual search and recognition, Vision Research, Vol. 50, No. 14, pp. 1338-1352, Jun 2010.

L. Elazary, L. Itti, Framework and implementation for perception, In: Proc. Vision Science Society Annual Meeting (VSS10), May 2010.

N. Noori, L. Itti, Symbolic Simulation: a grounded mechanistic account for processing symbolic information, In: Proc. 44th Annual Meeting of the Society for Mathematical Psychology (MathPsych 2011), Jul 2011.

N. Noori, L. Itti, Symbolic Simulation: a neural account for algorithmic and controlled information processing in human brain, In: Proc. 44th Annual Meeting of the Society for Mathematical Psychology (MathPsych 2011), Jul 2011.

N. Noori, L. Itti, Modeling forward and backward serial recall using a spatial registry assumption, In: Proc. Conference on Cognitive Science (CogSci 2011), Jul 2011.

N. Noori, L. Itti, Spatial Registry Model: Towards a Grounded Account for Executive Attention, In: Proc. Conference on Cognitive Science (CogSci 2011), pp. 1-6, Jul 2011.

N. Noori, L. Itti, Eye-Movement Signatures of Abstract Mental Tasks, In: Proc. European Conference on Cognitive Science (EuroCogSci 2011), (B. Kokinov, A. Karmiloff-Smith, N. J. Nersessian Ed.), pp. 110:1-110:6, May 2011.

N. Noori, L. Itti, Visuospatial attention shifts during non-visual mental tasks, In: Proc. Vision Science Society Annual Meeting (VSS11), May 2011.

*** Additional publications for applications of this project's research, where Dr. Itti's salary was supported by the project:**

Z. Li, L. Itti, Saliency and Gist Features for Target Detection in Satellite Images, IEEE Transactions on Image Processing, Vol. 20, No. 7, pp. 2017-2029, 2011. [2009 impact factor: 2.848]

Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image and Vision Computing, Vol. 29, No. 1, pp. 1-14, Jan 2011. [2009 Impact Factor: 1.474]

F. Baluch, L. Itti, Training top-down attention improves performance on a triple conjunction search task, PLoS One, Vol. 5, p. e9127, Feb 2010. [2009 impact factor: 4.351]

N. Parikh, L. Itti, J. Weiland, Saliency-based image processing for retinal prostheses, Journal of Neural Engineering, Vol. 7, pp. 1-10, Jan 2010. [2008 Impact Factor: 2.737]

C. Siagian, C.-K. Chang, R. Voorhies, L. Itti, Beobot 2.0: Cluster Architecture for Mobile Robotics, Journal of Field Robotics, Vol. 28, No. 2, pp. 278-302, March/April 2011. [2010 2-year impact factor: 3.580]

*** Review papers:**

F. Baluch, L. Itti, Mechanisms of Top-Down Attention, Trends in Neurosciences, Vol. 34, pp. 210-224, March 2011. [2010 impact factor: 13.320]

A. Borji, L. Itti, State-of-the-art in Visual Attention Modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, In press. [2011 impact factor 5.027]

# New Bayesian framework for perception

Student supported: Lior Elazary

## Abstract

A biologically-inspired framework for perception is proposed and implemented, which helps guide the systematic development of machine vision algorithms and methods. The core is a hierarchical Bayesian inference system. Hypotheses about objects in a visual scene are generated "bottom-up" from sensor data. These hypotheses are refined and validated "top-down" when complex objects, hypothesized at higher levels, impose new feature and location priors on the component parts of these objects at lower levels. To efficiently implement the framework, an important new contribution is to systematically utilize the concept of bottom-up saliency maps to narrow down the space of hypotheses. In addition, we let the system hallucinate top-down (manufacture its own data) at low levels given high-level hypotheses, to overcome missing data, ambiguities and noise. The implemented system is tested against images of real scenes containing simple 2D objects against various backgrounds. The system correctly recognizes the objects in 98.71% of 621 video frames, as compared to SIFT which achieves 38.00%.

## Introduction

In the 1970's and 80's many researchers believed that we would soon be able to achieve machines that can think and act for themselves. Although a number of impressive algorithms have emerged to efficiently solve many artificial intelligence problems, these machines never materialized due to what we think is a failure in perception. For example, a computer can beat any known human in chess if the computer knows exactly where each of the pieces are on the board. However, given a camera, inferring the position of the chess pieces remains very difficult. For that reason, we believe that if the perception problem were solved, many of the machines promised in the past would be able to materialize. Therefore, we propose a framework for perception, which can help guide the development of algorithms and methods in a more systematic manner. This will enable researchers to start concentrating on particular problems in perception whether being the structure of the system, or efficient computations when matching against a large amount of data or model representations. Additionally, we present a specific implementation for a visual perception in the realm of object recognition.

Using a visual sensor to recognition object has long been a problem in computer vision. Many algorithms have been developed, but they all seem to fall shy of good performance. This is often due to the fact that extracting low level features from a visual camera is often difficult and prone to errors. For example, edge detectors often have problems with noise edges, occlusion, missing edges and returning edges which do not belong the the object like specular edges (figure 1).

Figure 1: Example of an image taken with a camera, and the results of a Sobel edge detector.

Humans seems to have developed a solution to this problem which has remained a mystery so far. By examining visual illusions, we could attempt and extract some of the processes that are occurring in the brain. Looking at figure 2, we can see that the squares marked A and B perceptually seem to have a different shades of gray, where in reality they have the same shade of gray (pixel grey level value; you can check this by occluding the rest of the display with a piece of paper except for a small area around A and B). In this example, it seems that the brain is using high level information about the scene and shadows, to "fix" the low level information about values of the squares.

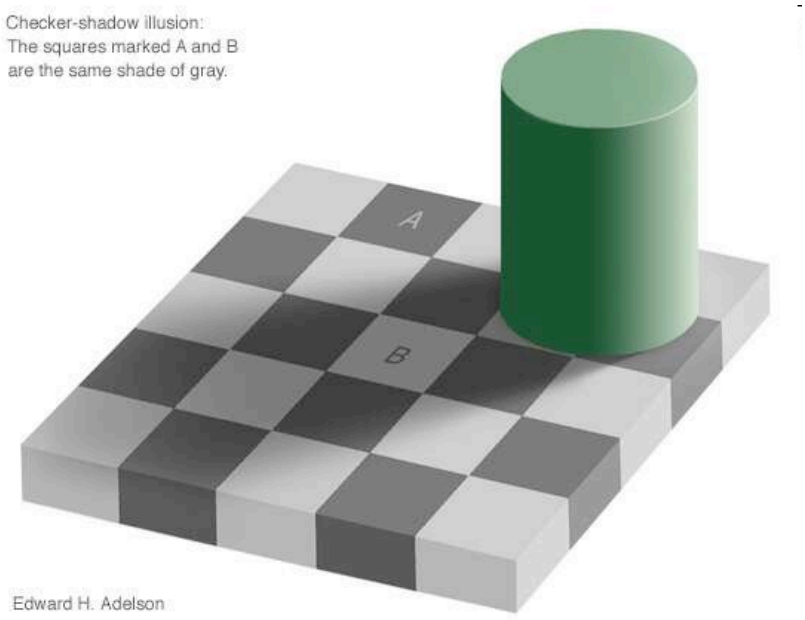



Figure 2: Checker-shadow illusion by Adelson.

Another visual illusion illustrating “Illusory Contours” can be seen in figure 3. Many people report a white triangle on top of a other triangle, even though there are no contrast changes on the white triangle. In this image, the brain has chosen to hallucinate the edges, even though they do not exist.

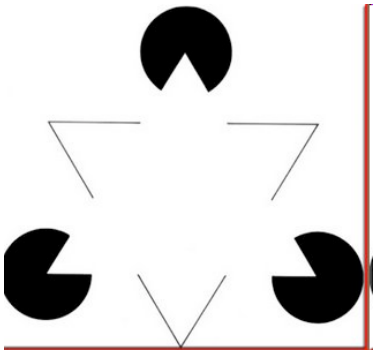


Figure 3: Kanizsa triangle (1955).

These illusions illustrate that contrast plays an important role in perception. Figure 4 shows visual stimuli in which ambiguity occurs due to missing or noisy local features. The images show how context (or prior information) can help resolve these ambiguities. The top left figure shows an example of a much localized use of context, where the center squares will be classified differently despite the fact that they have the same contrast. This is a well know visual phenomenon, which is thought to help in determining the true intensity of various patches. This phenomenon results from the interaction of neurons in the eye between the centers and surrounding regions. As a result, the different surrounding context changes the classification of the center. Another example of context use is depicted in the top right figure, where the circled image blobs are exactly the same (except for orientation) but get classified differently based on their placement in the scene. Lastly, the bottom center figure shows an ambiguity between the letters H and A where only context can help determine the true identity of these features.

In this work we use these illustrations to help and “fix” the low level features and provide more robust object recognition. In particular, we use the concept of hallucinations based on priors to fill in the missing data, which results in better recognition. We state that vision is nothing but controlled hallucinations, since most of the time we use the system to correct and fix low level features. To achieve this we use bottom-up features as proposals and not the data, and attempt to not make any major decision during the bottom-up feature extract. We then use top down knowledge to provide the decisions and impose priors on these features.

Figure 4: Example visual stimuli in which ambiguity occurs due to missing or noisy local features. Top left, the center squares will be classified differently despite the fact that they have the same contrast. Top right, the circled image blobs are exactly the same (except for orientation), but get classified differently based on their placement in the scene (obtained from (Torralba 2003)). Bottom center, ambiguity between the letters H and A.

# Problem Formulation

Figure 5: Example scene

Figure 5 depicts a typical scene were a typical percept of the scene could be the location and existence of cars. Previous approaches to solve this problem have been to build a model of a car, and use a sliding window approach to search at every location for the car. However, this might not be the most efficient manner to search through many large images. Mapping this approach to the MCMC framework, we can see how this can be improved. In this problem our goal is to determine p(W ) where W represents only two aspects of the world: the existence of a car and its position. In other words, we wish to model p(Car = T rue/False, Position = x, y). Note that the position is in visual coordinates, but we may wish to find the position in 3D. Using Bayesian mathematics we can formulate the probability of the world W given an image I in terms of the likelihood model p(I|W ) and the prior information p(W).

$$W \backsim p(W|I) \propto p(I|W)p(W)$$

However, p(W |I) might have an infinite solution space with complex structure, which would make it very difficult to solve in closed form. In the MCMC paradigm the system would generate a hypothesis for a particular state of the world (hypothesize that there is a car at x=5, y=6) and test it with a model of a car using the likelihood model p(I|W ). If the posterior probability is modeled as a state machine, where the probably of being in the current state is only dependent on the previous state, then this is know as a Markov Chain. In particular:

$$P(w_t|I_t) \propto P(w_t|I_t) \int p(W_t|W_{(t-1)})\, p(I_{(t-1)}|I_{(t-1)})\, dW$$

Particle filtering, histograms, or other distribution models can then be used to estimate the posterior density function. This is done by drawing samples from the distribution at strategic locations (i.e. the proposal distribution Q(x)). The proposal distribution can be determined from previous knowledge embedded in the system. One can therefore see that generating the hypotheses with a sliding window approach amounts to searching every possible solution in a discrete space, while using MCMC only samples the search space in specific places. Unfortunately, this MCMC can become very inefficient if we want to consider multiple objects, their full pose (x,y,z, scale, orientation) as well as other attributes like color and texture. As a result, a better sampling technique will need to be utilized.

Using the Metropolis-Hastings (MH) algorithm, we can build better proposal distributions which can be used for sampling based on prior information. The algorithm uses a proposal density Q(W′; Wt) which depends on the current state Wt to generate a new proposal W′ . The proposal is then accepted if:

$$U(0,1) < \frac{P(W')Q(W_t; W')}{P(W_t)Q(W'; W_t)}$$

Where U(0,1) is a uniform distribution between 0 and 1, p(W ) is the probability of our model (that is, was there a car there), while Q(W ′ ; Wt ) is the proposal distribution. For example, if we know that cars are usually on the ground, a proposal distribution can be: W=N(1/4 image height, 1/2 image height); where N is the normal distribution with a mean and variance. This would then choose locations which are closer to the ground and test them for the existence of cars. Note that if the car appeared in the sky, the system will take longer to find the solution.

Although the above method of using prior information would yield a faster convergence, it can still be improved further by “listening” to the data in the image. This can be achieved by combining saliency maps with the priors to generate better hypotheses. A simpler method of this has been proposed by Zhu et al (2002). and is known as data driven Markov chain Monte Carlo (DDMCMC). As a result, not only locations around ground level will be searched, but more concentration will be applied on regions which are more salient. Equation 1.4 shows the modification to the proposal distribution to include information from the sensor I.

$$Q(W_t; W'|I) \, aprox \, \frac{P(W'|I)}{P(W|I)}$$

The DDMCMC method have been successfully used in the past for image segmentation and scene understanding (Zhu et al, 2003) and can deal with noise and ambiguity. However, DDMCMC does not deal well with missing data and is still very computationally intensive for large hierarchical implementations of the framework. Fortunately, research in robotic navigation has been able to advance the use of particle filtering and geometrical constraints to combat some of these problems.

**Visual Perception Framework**

The framework is built from multiple hierarchical Bayesian inference modules that send and receive information to/from each other. Each module has two types of inputs and outputs: Data which is passed from lower levels in the hierarchy up, and Priors which are passed from higher levels down (Figure 6). Note that the Data/Prior can consist of multiple heterogenous streams. For example, a module responsible for finding square contours in an image will accept Data as corners and edges. It would then compute the probability distribution of squares existing in particular locations, compute the prior probability that the corners/edges should exist in particular locations, and bias the underlying edge and corner modules to correct or hallucinate any missing or noisy edges or corners.

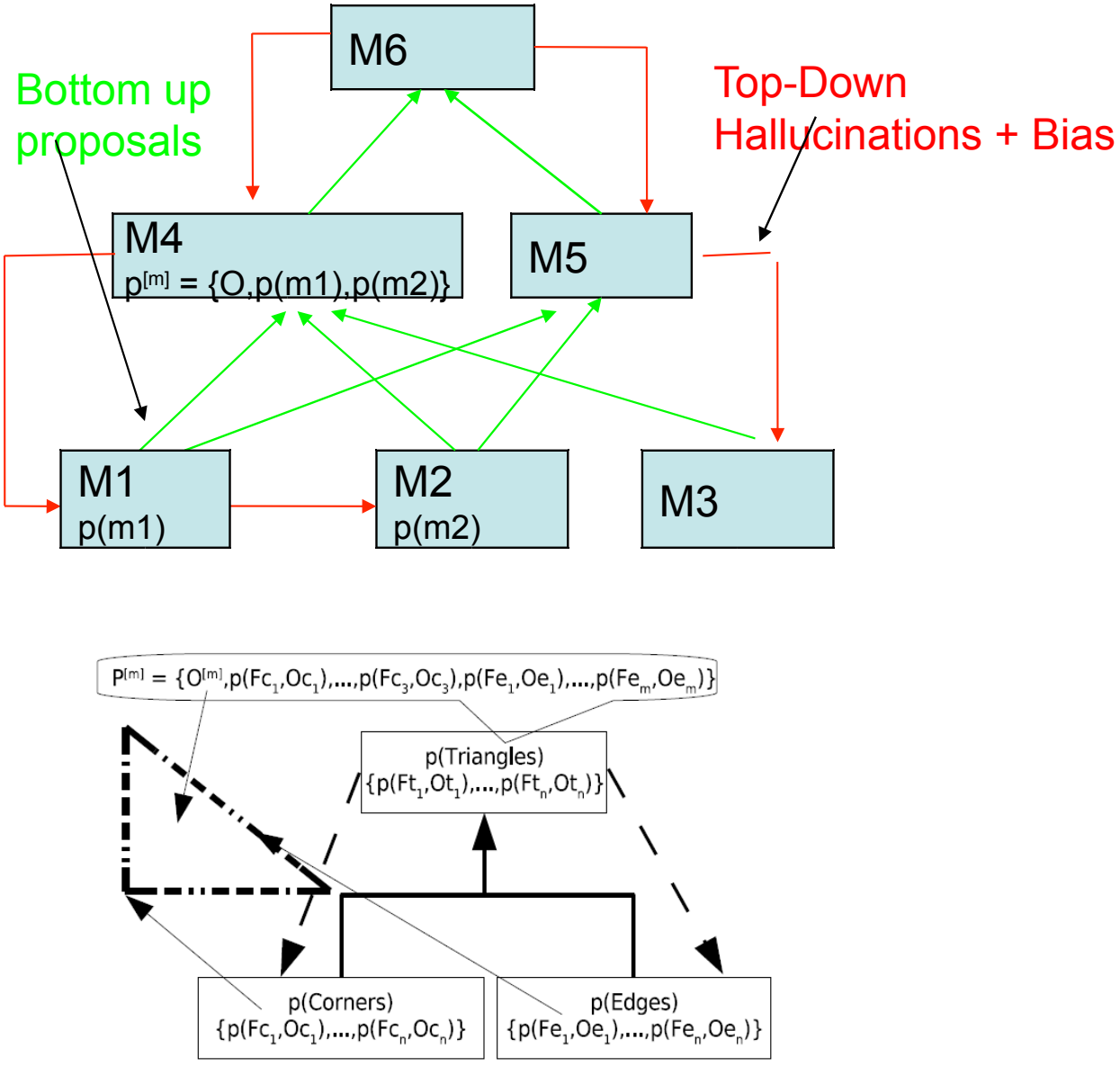



Figure 6: The basic hierarchy setup of modules (top) where solid line arrows represent feed forward connections Data and dashed arrows represent feedback Prior. An example triangle object where corners and edges are the features Fc and Fe, the pose of the triangle is Ot which corresponds to the center of the triangle can be seen in the particle P[m] (further described in the main text).

Each module is responsible for a particular belief within the system p(B|D), where B is the belief in some perception and D is the evidence for that perception. For example, the edges module would maintain the belief of edges existing in the world given the values from edge detectors, such as Gabor filters. This belief is computed using Bayesian inference with the ability to hallucinate data in a controlled fashion. Here we define hallucination as the ability to manufacture data that is not there. This is different from the view of priors, since we let the hallucinations offset the Data rather than scaling it, which results in hypotheses having positive values even if the underlying data has zero values. This is important, since often a particular module could maintain that there is no evidence (e.g., for an edge in a particular location) but a higher-level module would insist on these edges being there. For an example of this type of behavior in humans see the famous illusory contours (Kanizsa. 1955). However, if the system was allowed to hallucinate without control, it could always manufacture Data to satisfy its beliefs. To prevent this we use the concept of surprise (Itti and Baldi 2006), which is the KL difference between the prior and the posterior, to break away from "false" hallucinations.

In what follows we focus on a particular module that is characterized by a set of features F with a pose O, and which interacts with both lower-level modules (subscript L) and higher-level modules (subscript H). We thus define a particular belief in perception F = {O, FL } in the module of interest to consist of a pose O and a subset of features FL coming from lower-level modules (e.g., O might be the pose of a triangle, FL would be its three corners, and F would be used by higher modules as evidence for a particular triangle in a particular pose; Figure 6). That is, each module models a centralized relationship O for a set of lower-level features FL and maintains it as F . Each module then computes the posterior p(F | L ). In visual perception the pose O can consist of 2D or 3D pose parameters, while the features F can be edges, corners, color, motion, elementary shapes, complex objects, etc. Therefore the pose O can be seen as a constraint on the arrangement of a set of features FL . The posterior is calculated based on the observation (Murphy) that the features are independent given the pose. This insight allows the posterior to be factored exactly as:

$$P(F_{1:t}|F_{L,1:t}) = p(O_{1:t}|F_{L,1:t}) \prod p(F^i_{L,t}|O_{1:t}, F_{L,1:t})$$

where 1:t represent all data obtained from time 1 to time t (F1:t = F1 , F2 , ..., Ft ) while i = 1 · · · n is the particular feature in the model. For example, a model of a triangle that consist of 3 corners would have a pose O (e.g., location of center of mass, rotation angle, scale factor) and a set of 3 features FL,t for each corner in relation to O. The independence assumption makes it possible to estimate the pose O as well as estimating the probability of each feature FL . Where:

$$p(F^i_{L,t}|O_{1:t}, F_{L,1:t}) = \eta\, p(F_{L,t}|O_t, F^i_{L,t})\, p(F^i_{L,t}|O_{1:t-1}, F_{L,1:t-1})$$

Here η is a constant which corresponds to the normalizing factor (derived from Bayes' equation). p(FL,t |O1:t , FL,1:t ) is then used as a feedback to supply a prior to the lower-level modules. This prior is then added to the likelihood that the module maintains p(FL,t |Ot , FL,t ). Therefore our specific module of interest would obtain its prior from higher module(s) H as:

$$p(F_{L,t}|O_t, F^i_{L,t}) = p(F_{L,t}|O_t, F^i_{L,t}) + p(F^i_{H,t}|O_{H,1:t}, F_{H,1:t})$$

where OH and FH are the prior pose and feature coming from a higher-level module. We use a non linear saturation to cap p(FL,t |Ot , FL,t ) at 1 if its value is greater then 1. This can be interpreted as either taking the likelihood that the module maintains or the prior from a higher level to form the likelihood for the feature existing in a particular pose.

To ensure that the hallucinations do not produce too much false data, the module evaluates how surprising the addition of the data has been, as proposed in (Itti and Baldi 2006). This is achieved by taking the KL distance between the posterior p(Ft |FL,t ) and the full prior p(Ft ):

$$Surprise = KL(p(F_t|F_{L,t}), p(F_t))$$

If surprise surpasses a particular threshold θ, the likelihood p(FL,t |Ot , FL,t ) is reset to a uniform distribution. Future work will study how to only correct the wrong hypothesis or learn a better model for the module to decrease surprise in the future. Note that θ is the only parameter that needs to be tuned for each module. This parameter determines how quickly each of the modules adapts to changes and learns to better predict the world. For example, if a feature has moved from one position to another, a large surprise value will occur. However, if another module could have predicted the movement and shifted the probability distribution accordingly, then the system is not surprised. Therefore, learning such a predictive model would help explain the world with better hypotheses.

Particle filtering (Montemerlo et. al. 2002) is then used to efficiently compute and store the probability distributions within the modules, which allows the system to support multiple hypotheses. However, in some particular modules where the probability distributions can be estimated from a single Gaussian, the simpler Extended Kalman filter is employed (Maybeck 1979). Here we describe the particle filter approach due to length constraints. Each particle Pt (where m is one out of M particles) has an estimated pose Ot and i,[m] a set of estimated features p(F ) (refer to (Montemerlo 2002,2003) for more details on the process):

$$P_t^{[m]} = \{O_t^{[m]}, p(F_t^{1,[m]}), ..., p(F_t^{n,[m]})\}$$

In each iteration a new pose for each particle Ot is estimated using the features from previous layers as well as the priors. Bottom-up computations help guide the system toward probable “right” hypotheses. The inspiration comes from (Zhu et. al. 2000) and (Montemerlo et. al. 2003) and is known in the literature as Proposal Distributions. Therefore, the pose O in the module is sampled based on prior data along with the current feature observations FL from lower-level modules. This bottom-up approach is based on two types of computations, saliency maps (Itti et. al. 1998, Itti and Koch 2000) and the Generalized Hough Transform (Ballard 1987). The saliency maps are used to pick relevant features/locations for a particular module, which are then fed into a Generalized Hough Transform computation to estimate the pose as defined earlier. The saliency computations provide an efficient manner of disregarding information in the presence of clutter and noise, since less data needs to be processed. It is important to note that the full probability distribution is fed into the next layer and not just its maximum or other single value summarizing the data. Therefore, the bottom-up probabilities only guide the module to the correct pose/feature, but do not make a decision as proposed in other bottom-up approaches (Serre et. al. 2007). The reason for this is that a higher level in the hierarchy

might find that less probable data received in the lower level is more relevant for the current perception. The proposal distribution can be reformulated as follows:

$$P_t^{[m]} \sim p(O_t | O_{1:t-1}^{[m]}, F_{L,1:t}) = \eta^{[m]}\, p(F_{L,t} | O_t, O_{1:t-1}^{[m]}, F_{L,1:t-1})\, p(O_t | O_{t-1}^{[m]})$$

where η is a normalizing constant for particle m, $p(O_t | O_{t-1})$ is the prior distribution and

$$p(F_{L,t} | O_t, O_{1:t-1}^{[m]}, F_{L,1:t-1}) = \sum_{i=1}^{n} p(F_{L,t} | F_{L,t}^{i}, O_t)\, p(F_{L,t}^{i} | O_{t-1}^{[m]}, F_{L,1:t})$$

To evaluate $p(F_{L,t} | O_t , O_{1:t-1} , F_{L,1:t-1})$ we first use saliency computations to filter out irrelevant data. The saliency computation differs based on features. For example, the full saliency map of (Itti and Koch 2000, Itti et. al. 1998) can be used to filter out the background and only concentrate on objects, while the max- normalization of (Itti et. al. 1998) can be used to filter out edges that correspond to textures (repeating patterns), and only let extended contour edges pass through. This level can be seen as an attention level, which can also be biased based on a task as proposed in (Wolfe 1994, Navalpakkam and Itti 2007) to further reduce the amount of data. Once this step is completed, the Hough transform is used to estimate the pose based on a model. This enables an efficient running time of O(NM) where N is the number of features that passed though the saliency maps and M is the number of pose parameters to estimate. To sample from the Hough map each module first converts the map into a probability distribution by normalizing the map, then using the concept of inhibition of return (Itti et. al. 1998) picks local maxima for the sample pose.

The proposal distribution does not yet correspond to the posterior so we must resample to evaluate the hypotheses of the particles. To account for multimodal distributions, we use stratified sampling. The importance weights are calculated as follow:

$$w_t^{[m]} = \eta \prod_{i=1}^{n} p(F_{L,t}^{i} | O_{1:t-1}^{[m]}, F_{L,1:t-1})$$

where $p(F_b | O_{1:t-1} , F_{L,1:t-1})$ is a known model which maps between feature positions and poses. Note that currently we assume that models of the features are given for particular objects, and the system does not learn models but just evaluates them. Therefore, the feature positions within a model are never updated since they are known precisely. Future versions of the framework will incorporate learning for new objects, which will include updating the feature positions and their model.

# Implementation of the framework for visual perception (Version 1.0)

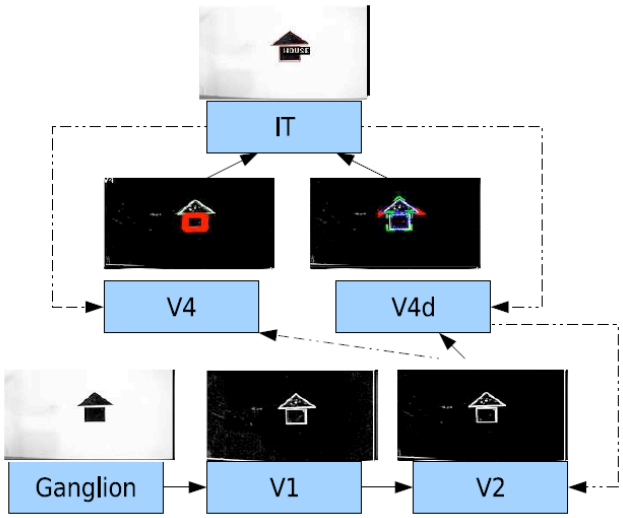



Figure 7: The implementation of 2D perception. See text for details.

An implementation of the framework was developed for visual perception. The modules are inspired from the visual structures proposed by (Marr 1982 and Hummel and Biederman 1992) and various visual cortex regions and their properties. The world for this model consists of shapes in a 2D world subjected to translations and rotations as well as cluttered backgrounds. Figure 7 shows the modules and their feedback.

The ganglion cells modules are responsible for p(Wi |I) which computes the belief over luminance at a given location Wi in the world given image data I. This image data comes from the luminance value of a camera. Since this probability distribution can be modeled by a Gaussian, this module uses the Kalman Filter equations for the update, where the prior comes from the V1 cells layer.

The next layer in the system computes the orientation and location of edges and is named V1. This is achieved by using the Sobel operator (Feldman 1968) as a saliency operator to give evidence for edges. Since the Sobel operator can estimate the pose from the filter directly, the Hough transform is skipped for this level. To model edges, Gabor filters are used. Given the 2D nature of the input, only the position and orientation are used for the pose. Since a given location can respond to multiple edges (e.g., corners) the particle filter approach is used. This module is an example of how efficient the system is at computing edges, since computing this space with just Gabor filters can be very time consuming in a serial process. However, the prior for this module comes from V2, which helps guide and improve the hypothesis of edges in this layer. This is also biologically plausible, since the tuning properties in primate V1 cortex just following onset of a stimulus ($\infty$ 40 ms) correspond to edge orientations, spatial frequencies and colors, while at a later time ($\infty$ 100 ms) are more sensitive to edges that correspond to global properties in the scene (Lamme and Roelfsema 2000).

V1 then provides input into the V2 module, which computes various Gestalt features and non-accidental properties (Biederman 2000) as its bottom-up computations. This stems from responses of many V2 neurons in visual cortex corresponding to illusory contours, and figure-ground segregations that can be achieved with Gestalt laws (Qiu and Heydt 2005). This module then computes the probability that a local edge belongs to an extended contour. This filters out edges from textures or background. In our particular implementation of the model, the work by (Grigorescu et. al. 2003)is used to improve the probability of edges belonging to a contour based on their neighbors. Future implementation will include a suite of Gestalt laws. The prior for this module comes from V4d and V4, which provide evidence for the contours.

The V2 module provides information to V4d and V4, which compute simple geometric shapes like squares, triangles, circles, etc, that we define as Geons. These shapes are inspired by the Geon hypothesis that complex objects are broken down into simpler geometric shapes (Biederman 1987). V4d is responsible for computing the vertices of shapes (in the case of a circle, small arcs are computed), while V4 computes the outline of the contour. This type of tuning response has been found to be computed by the neurons of primate V4 cortex (Pasupathy and Connor 1999 and Gustavsen and Gallant 2003). The max-normalization method proposed in (Itti 1998) is used for saliency computations while the generalized Hough transform is used to find the pose for basic shapes: square, triangle and circle. The probability that a shape exists given a pose is evaluated by assuming that each vertex is located within a Gaussian PDF and that the edges are along a straight line from each vertex. The prior for this module comes from the IT level, which provides the probability that various Geons exists.

The last module in the system, IT, computes the probability of objects having particular poses. It receives its input from the V4 module, and no saliency computations are used. Instead, the V4 Geons are fed directly into a Generalized Hough transform map from which the proposal distributions are evaluated. It has been found in the framework that the higher the module in the hierarchy, the less saliency computations are needed since the amount of data that passes through is minimal. The probability of an object given a pose is evaluated by assuming that each V4 Geon is located within some Gaussian amount of uncertainty in both position and orientation.

# Results for system version 1.0

Figure 8: The objects used in testing. From left to right House, Woman, Hat, Man, Man with Hat.

The implemented system was tested by using an overhead camera focused on a scene containing simple objects. The objects where composed of simple cardboard cutouts of shapes: squares, triangles and circles, which were arranged into familiar shapes (Figure 8). Five objects were used in the tests and they are House, Man, Woman, Hat (defined as a triangle on top of a circle) and Man with Hat. These

objects where then subjected to various translation and rotations along with complex background changes. To compare against state of the art, we used the SIFT method proposed by (Lowe 2004). SIFT can be seen as an implementation of this framework by observing that the keypoint selections using scalespace extrema are analogues to our saliency computations, the evaluation of keypoints against a database is a non-parametric mapping function for evaluating hypotheses (the keypoint descriptors in this case), and the generalized Hough transform is the final output of the system (an additional layer with no validation). SIFT was trained on 15 views of each object (3 shots at 5 different poses). SIFT achieved 100% recognition rate on that training set, confirming correct operation of the SIFT implementation. The results for a test set of 621 images are in Table 1.

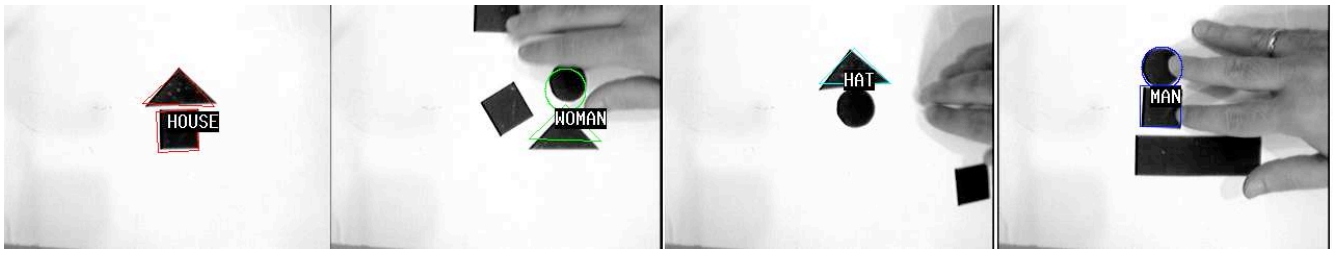



Figure 9: Example scenes tested in the model with clutter (hand, shadows).

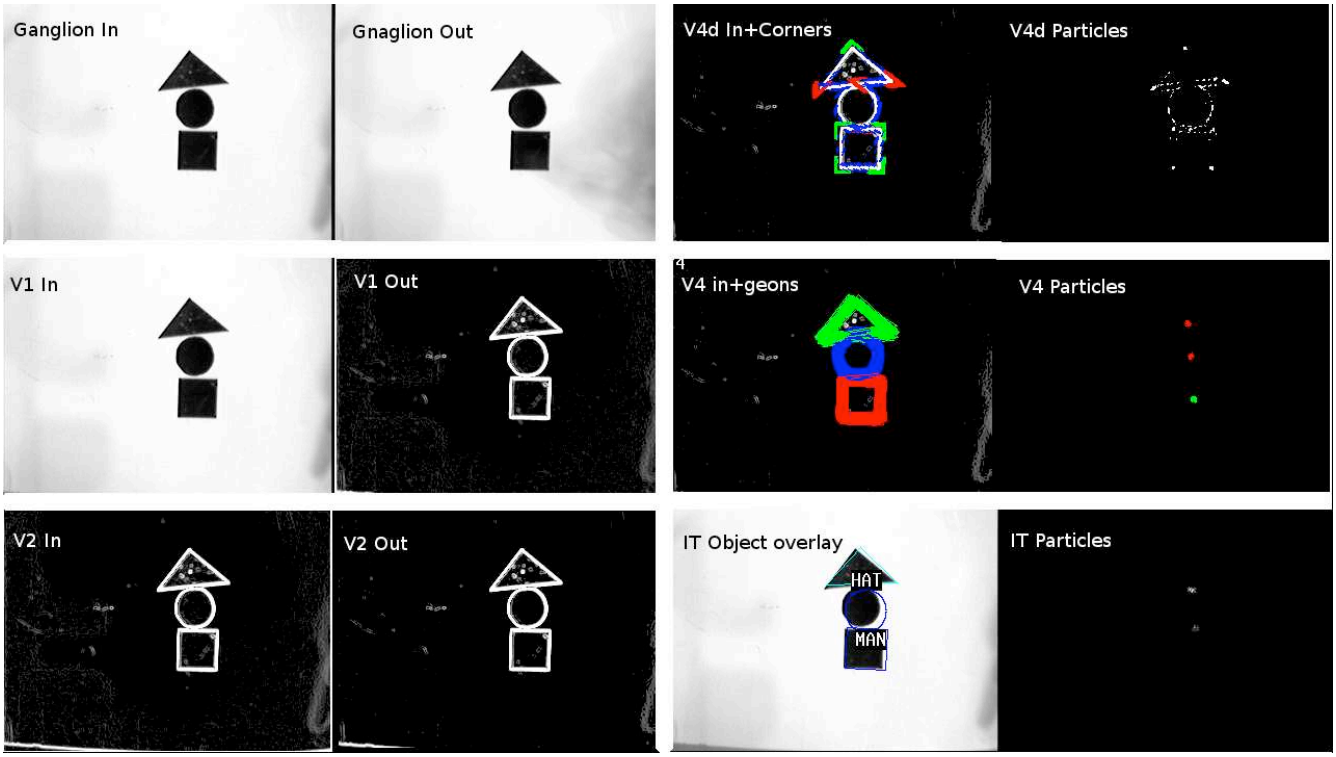



Figure 10: Input and Output maps for one scene in the implemented framework. From top left: Ganglion input (camera input) at left and cells output at right, middle left: V1 in/out, bottom left: V2 in/out, top right: V4d in + corners at left and particles at right, middle right: V4 in + Geons and particles at right, bottom right: IT with object overlay of at left and particles at right.

| Method | No clutter n=342 | Clutter n=279 | Total n=621 |
|---|---|---|---|
| SIFT | 41.81% | 33.33% | 38.00% |
| Proposed System | 100% | 97.13% | 98.71% |

Table 1: Results for the SIFT and the proposed system.

No temporal filtering was used during testing. This limited the system's ability to predict where an object might move, and to form better hypotheses (as in Isard and blake 1998). This was to determine system robustness with isolated static scenes. Adding temporal filtering should only improve recognition. Figure 10 shows an example scene with all the modules in the system, while Figure 9 shows just the IT output for some scenes. As can be seen the system correctly identifies the object. Examining the ganglion cell layer, most of the noise in the image has been eliminated. Looking at the V2 layer, most of the edges on the shadows have disappeared and only the outline of the contours remain. V4d shows the hypothesized corners and their positions without the feedback. Even though some of the corners have not been hypothesized correctly, the system is still able to correctly identify the objects in the scene. V4 shows how the correct Geons are hypothesized and the particle positions, while IT shows the correctly identified objects .

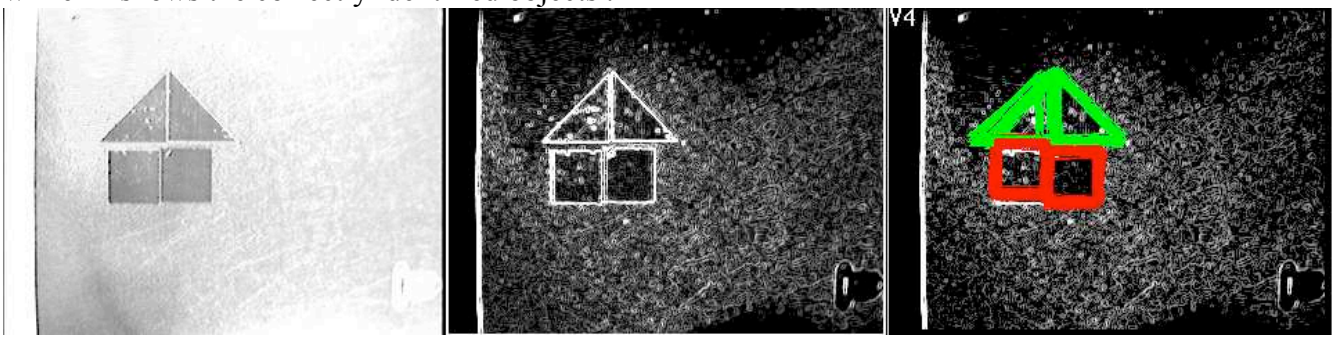



Figure 11: The responses of V1 V2 and V4 to a two house scene with a complex background. From left to right: input image, V1 response, V2 response with V4 response overlaid on top.

Figure 11 shows a selected complex background with the responses of V1, V2, and V4. As can be seen there are initially many false edges that the system needs to consider. Just using a standard Hough transform would result in many false positives. However, V2 already cuts down on a lot of the edges, while the verification step in V4 is able to narrow down the correct hypothesis.

# Implementation of the framework for visual perception (Version 2.0)

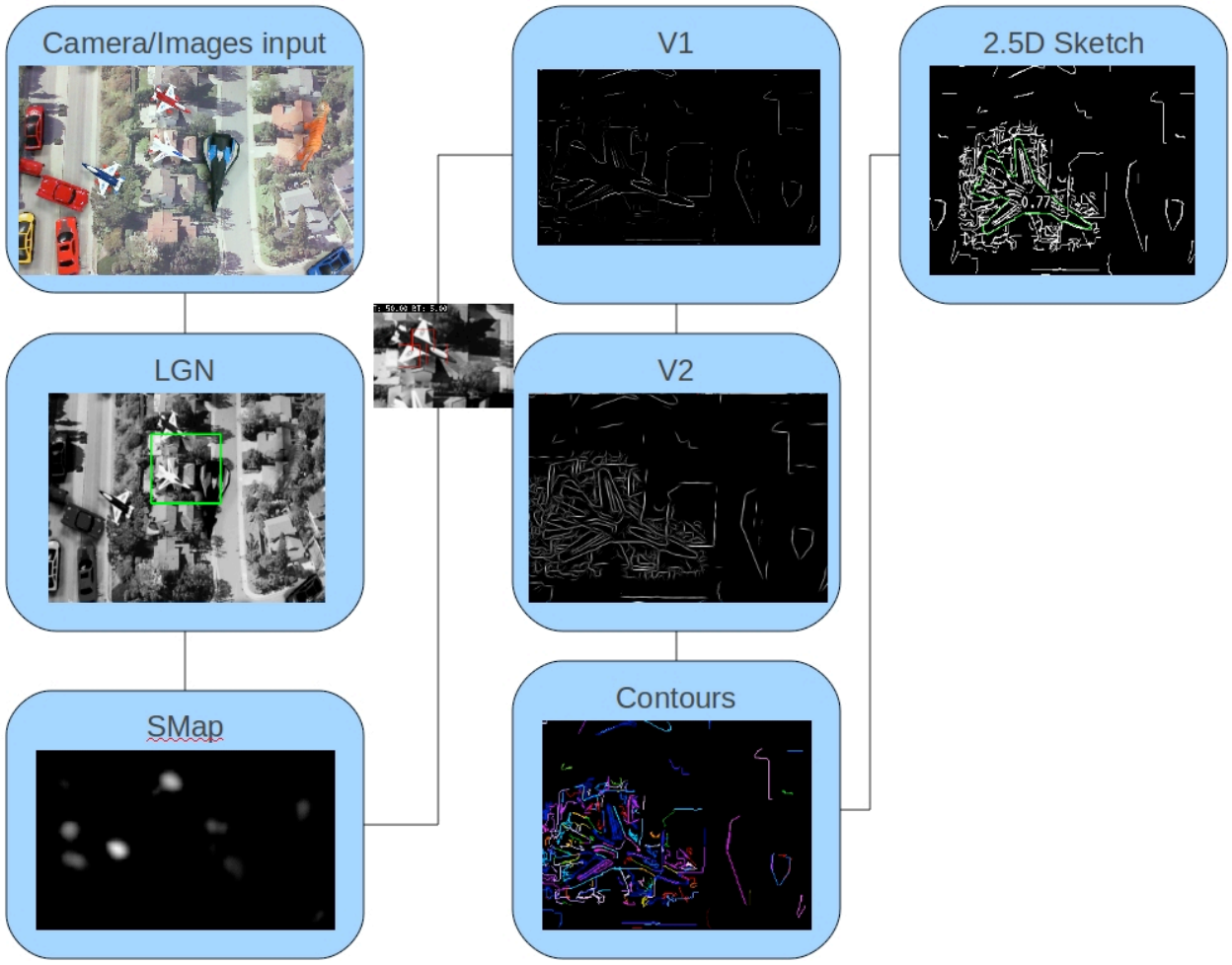



Figure 12: System version 2.0 modules. See text for details.

The current framework has been expanded and improved with a few more modules. Figure 12. shows the current blocks implemented. The module V2 now implements the tensor voting framework proposed by G. Medioni (2004). This boosts the edges and makes them more salient if they are continuous. The Contours module implements a simple edge following algorithm which prefers edges with the same orientation, magnitude and distance. The contours are then approximated with lines. Lastly the 2.5D Sketch module provides 2D shape recognition with similarity transformation. This is done with a Generalized Hough transform to propose shapes. The shapes are then evaluated using a directional chamfer matching for efficiency.

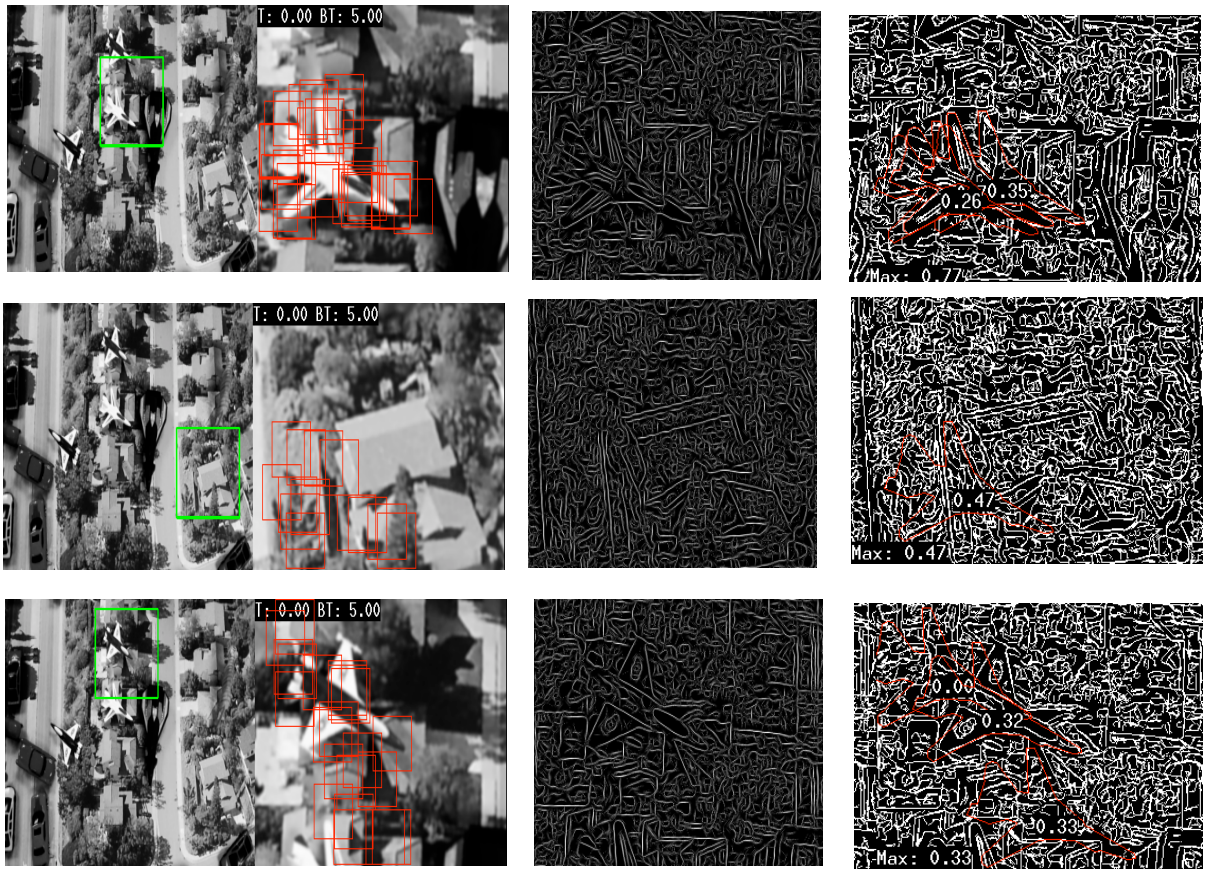



Figure 13: The results of the output of V1 for various locations in the image with a 0 threshold on the edge detection (mark any small magnitude edge as an edge. The left image shows the full image with the attended 320x240 image. The middle image show the edge detection, while the right image shows the proposals shapes along with their corresponding probabilities.

We also explored a method for tuning the various parameters in the system using the notion of bottom-up and top-down processes. This can be seen as biasing various detectors to give different results. One classic example is choosing a threshold for edge detectors (Figure 13). The system was developed to explore the parameter space using the current framework. This was done by setting the initial threshold value to a high number, which reduced the amount of data to the system. For example, setting the edge threshold in V1 produced less edges available to V2. The system will them propose a shape, evaluate a probability for it, and send a bias signal to reduce the threshold values in regions where there are no edges. This results in two outcomes. Either the shape is found, in which case the probability will increase due to the additional correct edges, or the shape is not found which will reduce or keep the probability the same (we determine that a shape is found if more the 75% of the edges contribute to the match). The process will then continue as long as the probability is increasing, or stop if its not. Even though this is an iterative process, about 3 iteration on average is required to extract the shape.

Lastly, we improved the efficiency of the matching process using Fourier Descriptors as top down features. The Fourier Descriptors have been used in the past for shape analysis (Charles et. al. 1972,

Otterloo 1991), character recognition(Persoon and Fu 1977, Rauber 1994), shape coding ( Chellappa and Bagdazian 1984), shape classification ( Kauppinen et. al. 1995) and shape retrieval (Guojun and Sajjanhar 1999, Sajjanhar 1997, Huang and Huang 1998). Fourier Descriptors are a method of extracting a shape signature in a form of a vector. These signatures are often invariant to scale, rotation and translation, which makes them very efficient at matching shapes.

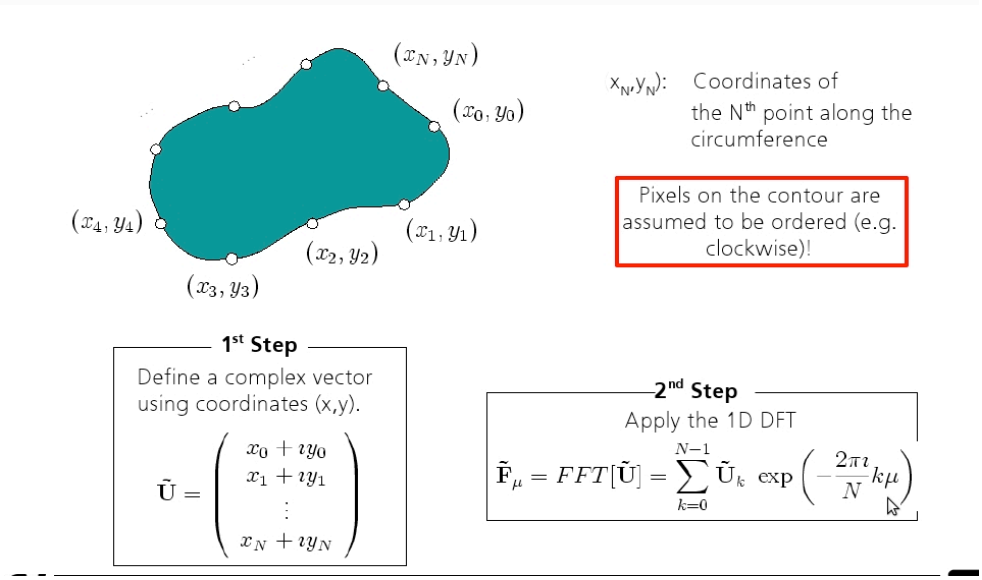



Figure 14: Fourier Descriptors formulation.

Figure 14 shows how the co-ordinates of a boundary are collected into a complex vector (1st step). The 2nd step assigns the Fourier descriptor by essential giving a spatial frequency that fit the boundary points. The first Fourier component (the d.c. Component) is simply the mean value of the x and y (I.e the center of the boundary). The second component gives the radius of a circle which best fits the points. Lastly, higher order components give more details on the contour with details occupying the higher frequencies (Figure 15).

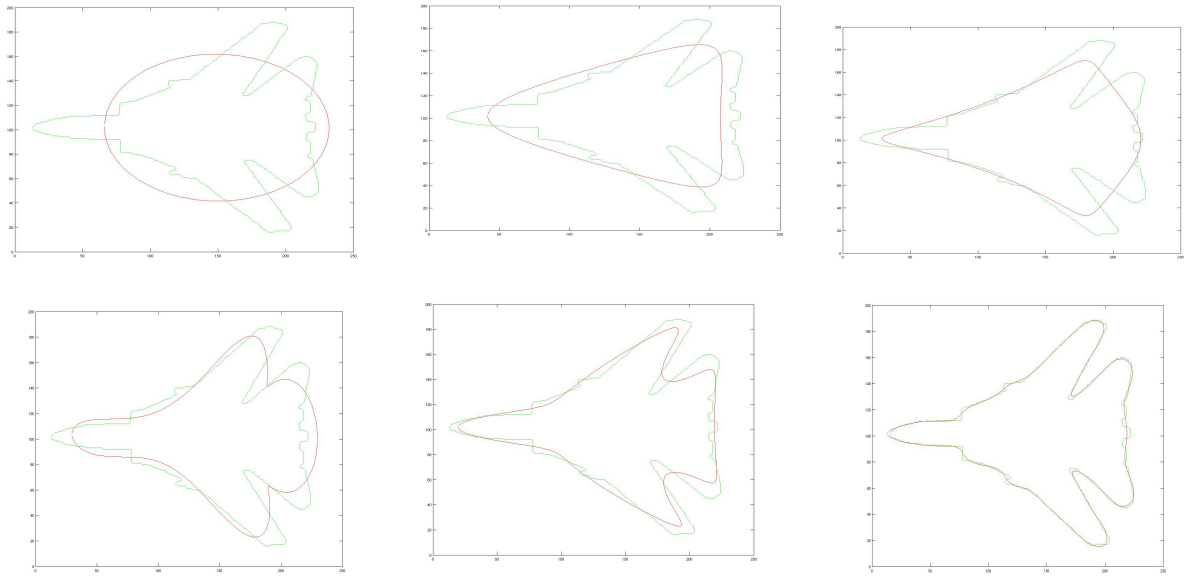


Figure 15: Fourier descriptor components for an airplane contour. Top row from left,1$^{th}$ 2th and 3th components. Bottom row, 5$^{th}$ 6$^{th}$ and the 20$^{th.}$

**Results for system version 2.0**

The first simple test consisted of a toy world filmed with a camera using a robotic arm. In the scene, a toy airplane was the target as seen in figure (13,16, and 17). Figure 13 shows an example where the threshold was set to 0. As can be seen, there are many edges to process, which took the system 15.62 seconds on average to extract and evaluate the shape proposals. On the other hand figure 16 and 17 show the iterative process for extracting the shape for the target object and a non target object. Each iteration took 1.17 seconds, which results in the shape being extracted within 3.5 seconds.

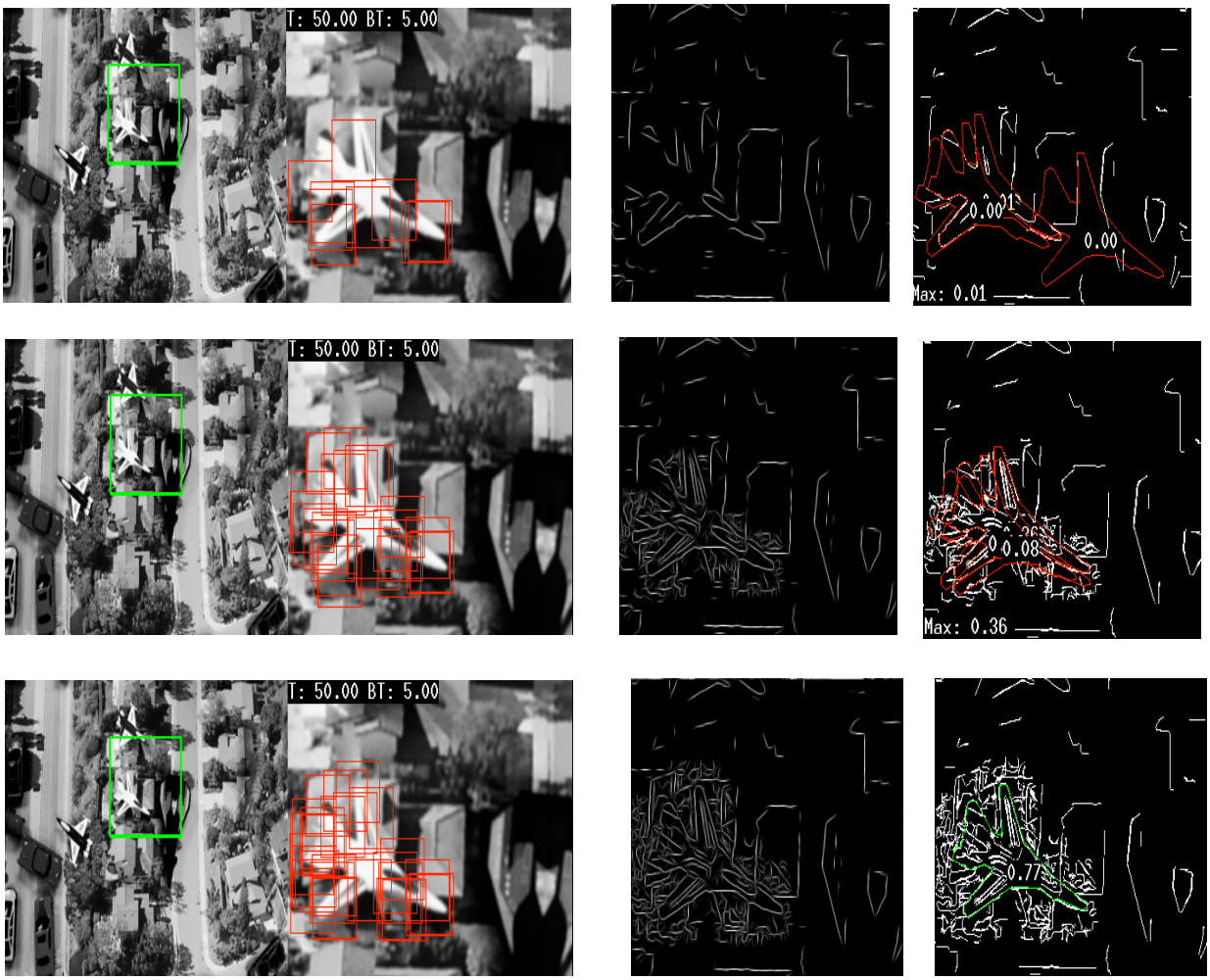



Figure 16: The iterative parameter tuning for edge selection. Three iterations are shown from top to bottom. The green box on the left indicated the attention location, while the red boxes indicate the biasing location. As can be seen, at each iteration the proposed shape improves the probability of the shape, which causes more edges to be biased. At the edge of the third iteration, the shape has a probability of more then 75%, which causes the system to stop and mark the shape.

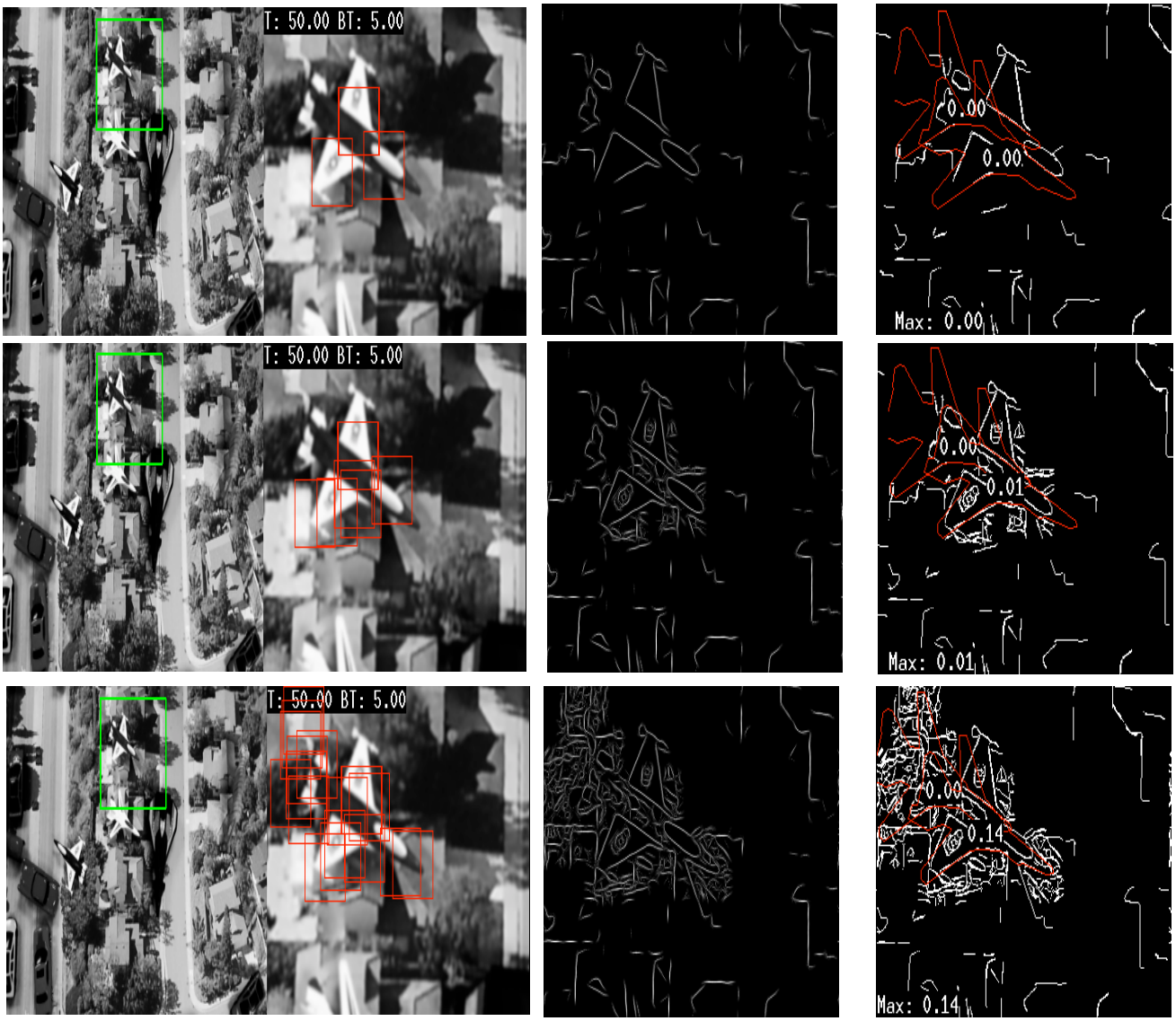



Figure 17: The iterative parameter tuning for edge selection. Three iterations are shown from top to bottom. The green box on the left indicated the attention location, while the red boxes indicate the biasing location. As can be seen, at each iteration the proposed shape improves the probability of the shape, which causes more edges to be biased. However, the shape never fully matches the target which causes subsequent iterations to not increase the probability, terminating the process.

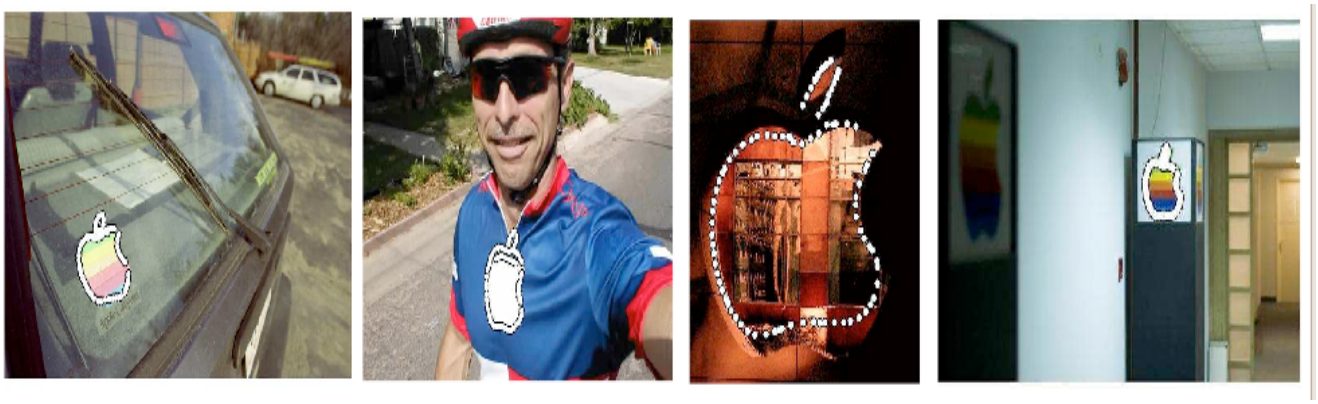


Figure 18: Example images from the ETHZ dataset.

The next set of tests were perform on the ETHZ dataset apple logos (figure 18) so it can be compared to the state of the art system. The dataset contains 255 test images depicting five diverse shape-based classes (apple logos, bottles, giraffes, mugs, and swans) in various scenes.

Figure 19 shows and example image with the output of the system before the dynamic biasing and after the dynamic biasing. As can be seen there are many false positives before the dynamic biasing, and only one after. This allows the system to lower the threshold on the matching and find more images with less false positives. Figure 20 shows the system being run with 30% threshold on the edges, 10% threshold and the dynamic biasing. As can be seen a 10% threshold has a lower detection rate then the 30% threshold. This is due to the fact that at 10% threshold there are less edges. However, at 30% threshold, we get more detection but also more false positive. Using the dynamic biasing, the system is able to take the best of both worlds. It uses the 30% to lower the false positive rate, and only lowers the threshold on probable candidates.  Lastly, figure 21 shows the results of the system in comparison with other methods. As can be seen, the dynamic biasing is able to have a greater detection rate with less false positives.

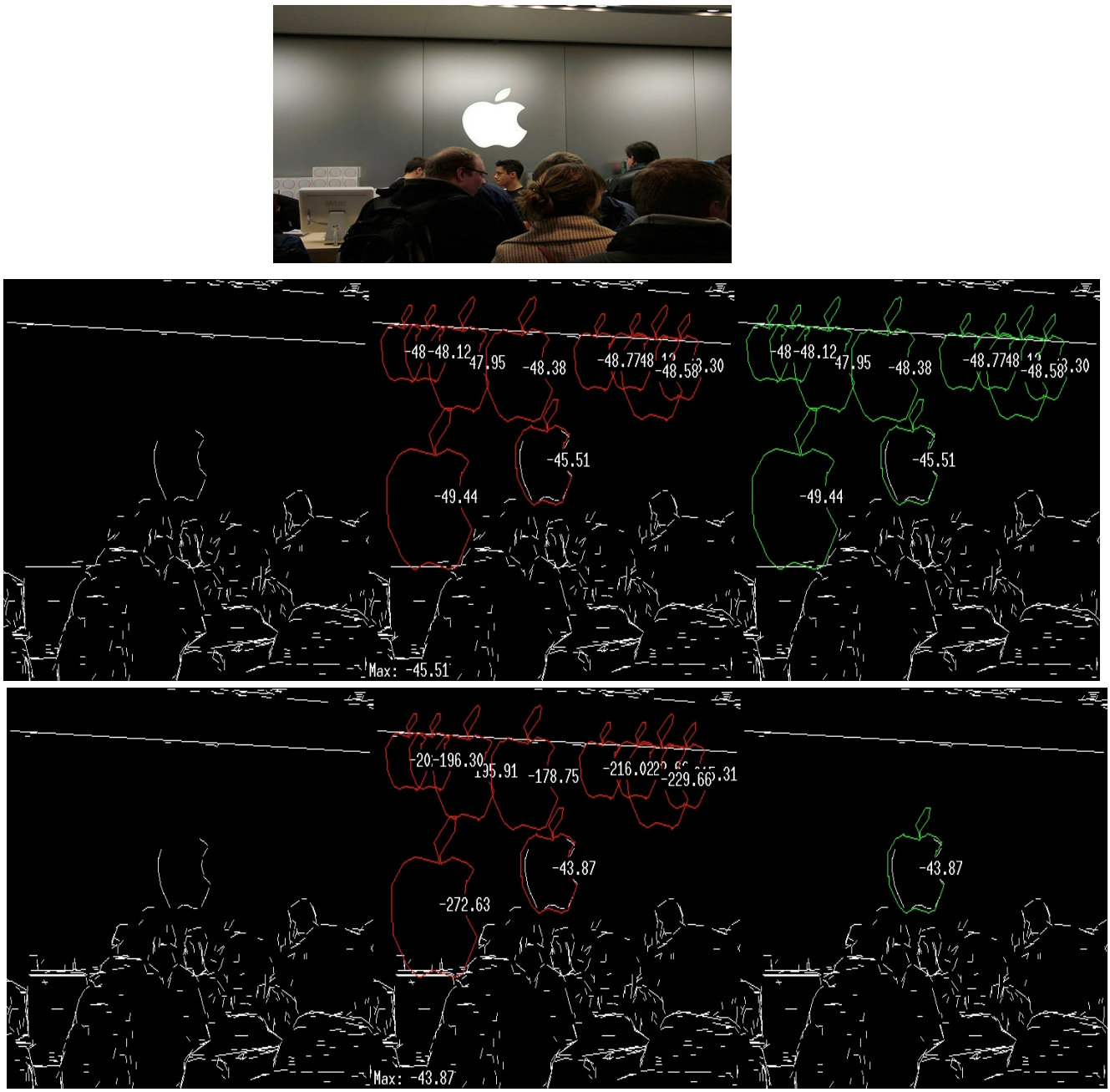



Figure 19: Example system output from a given image (top). The middle row represent the edges, input and output before dynamic biasing. The bottom row shows the system after the dynamic biasing.

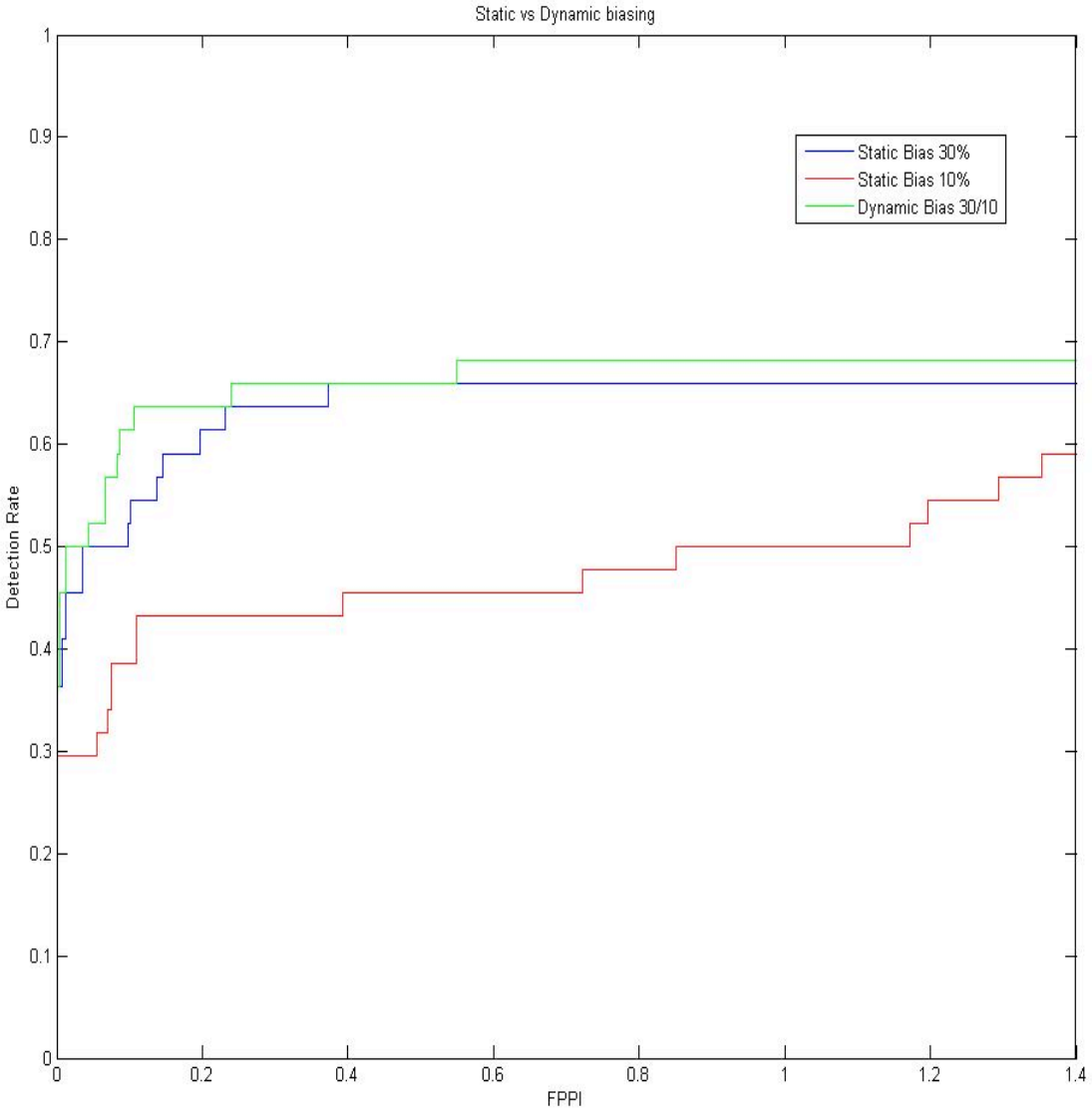



Figure 20: Running the system with fixed threshold and the dynamic biasing.

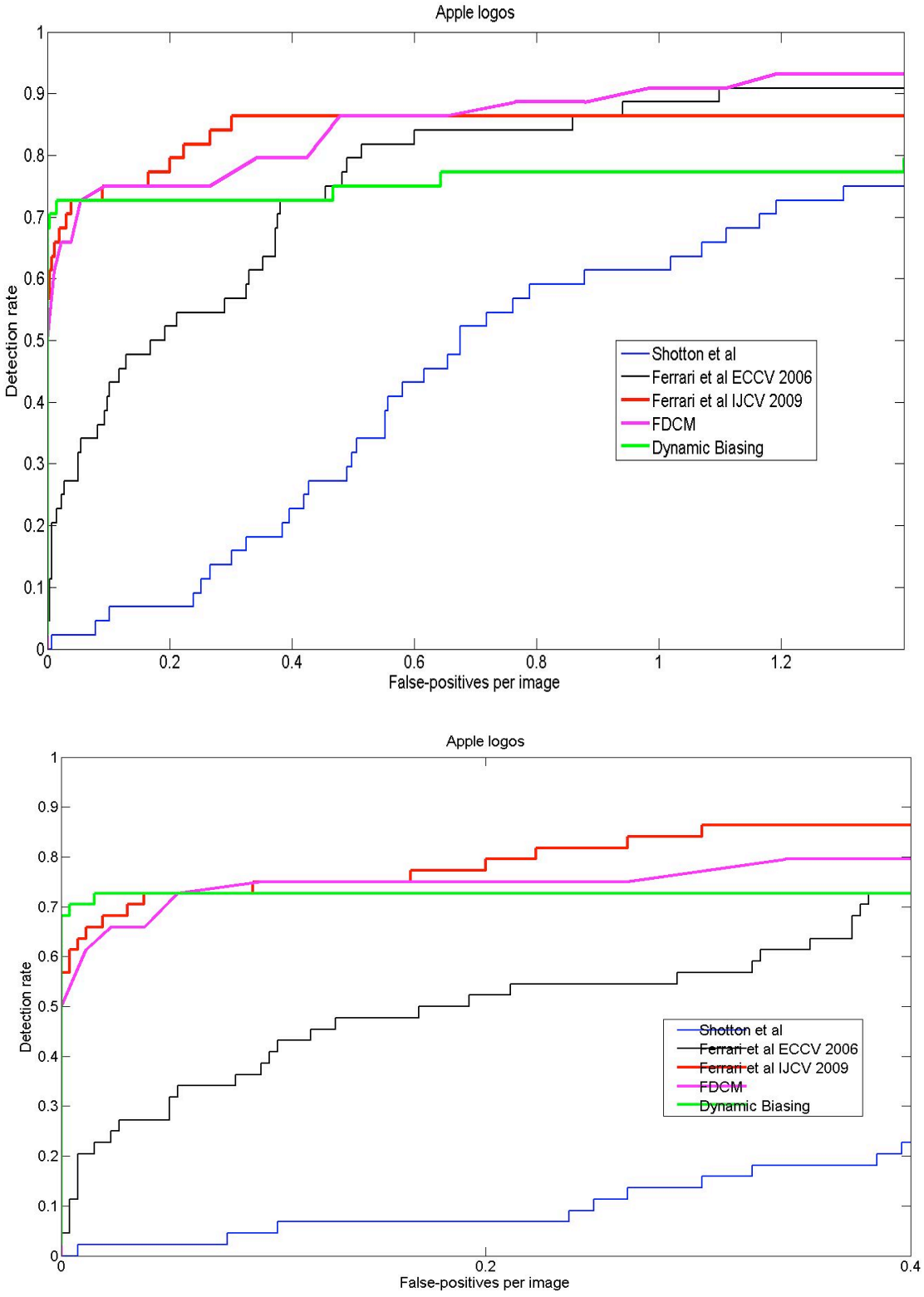



Figure 21: Results on apple logos comparing to other algorithms. Bottom image is a zoomed version of the top image.

# Conclusions

In this work, we propose a biologically-inspired Bayesian framework to solve the perception problem. Importantly, the framework is helpful in designing and analyzing algorithms for perception. For example, the structure of the modules proposed in this paper is based on a simplistic view of visual cortex. There are many other regions in visual cortex that can be utilized and tested. For example, adding an MT layer can help in forming fast and more correct hypotheses. The framework allows the addition of such module without having to change the system. We believe that this is a trait that is needed in a framework of this magnitude since the ultimate structure for perception would need to be determined by many researches. This would then allow some researchers to work on specific aspects of the problem and give them the ability to test how it fits together with other modules (or which modules would need to be changed). Additionally, the parameter tuning has shown a great improvement of efficiency as well as a reduction of false positives (phantom shapes). Currently, the generalized hough transform takes the majority of the time, as the system needs to scan over 4 parameters (2 for position, scale and rotation) for each shape.

In the future we are planning on taking the implemented system to a full 3D and adding more layers like shape from shading, texture, better Gestalt models, etc. Additionally, we plan on implementing learning for the framework so that we can handle more complex structures. For example, using the same methods as in the SLAM algorithms we can learn the probability of features existing for particular objects. Lastly, we plan on studying how actions can affect perception by biasing modules in the system to achieve perception that is relative to the current task. For example, playing video games requires the understanding of the scene, but there are many irrelevant details that do not need to be parsed or perceived in order to win the game.

# Neural basis of mental cognition

Supported student: Nader Noori

**Introduction**

Symbolic problem solving has proved to be powerful in solving practical problems ranging from counting one's finger to building intelligent agents. Cognition by means of abstract symbolic concepts in an algorithmic manner is one of the tenets of mathematical cognition. Identifying the relationship between this evolutionarily newly emerged symbolic machinery and rudimentary older modal systems has motivated numerous studies mostly focused on grounding representation of symbolic concepts. However recent evidences emerging from neuroimaging and patient studies suggest that modal systems for visually guiding actions in space play a role in mental operations on symbolic information that is beyond representation of symbolic concepts (Koenigs M. et al 2009, Knops A. et al. 2009).

Motivated by these findings we have been seeking a grounded mechanistic model for algorithmic controlled information processing in human brain. We studied the impacts of mental tasks with symbolic and presumably non-visual tasks on the visual system. We learned that mental tasks with symbolic content that with demand for active memory manipulation imposes a load on oculomotor system and impairs the visuospatial short term memory.

To explain our findings we propose a critical role for a spatially organized short-term memory which is used for anchoring task relevant items into the space. These anchors are used for selective processing of the maintained information. Selective processing of information (such as deletion of item from memory) in turn is made possible through shifts in spatial attention towards registry location of the item of interest in the space.

This registry system along with an articulatory system for hashing items into phonological codes, and a system for performing and monitoring sequential actions provide necessary mechanisms for employing overly-trained networks for processing limited set of activated items in arbitrary algorithms. We have evaluated our hypothesis by detecting process related traces of mental symbolic operations in both eye movements of human subjects and visuospatial short-term memory of objects in the environment.

**Eye tracking experiments**

We studied distributions of low amplitude gaze shifts made during a single mental sorting task in front of a blank screen and noticed that unlike active maintenance of a string of decimal digits, sorting them into order induces gaze shifts that carry information about the mental stimulus for the sorting (Figure 1.) . For example we noticed that strings of digits with flipped order of items result to symmetric distributions of gaze shifts (Figure 2.). We interpreted the presence of the information about the mental executive task in the oculomotor system as the result of the active involvement of the visuospatial system in the manipulation of memory items.

To give an account for the involvement of the visuospatial system in the process of memory manipulation we proposed the Spatial Registry Hypothesis (SRH) which assumes a functional role for brain regions with visual-spatial encoding features in registering memory items in a spatially-

organized short-term memory. We assume that a task-relevant items in the working memory may register with a corresponding visuospatial short-term memory (VSSTM). The spatial registry may occur when selective access to memory items is required. This access might be facilitated by shifting spatial attention in this internal registry space and this would give an account for the induced gaze shifts during the mental executive tasks.

This assumption is consistent with findings of neuroimaging and neuropsychological studies that have shown that the posterior region of the parietal lobe is critically engaged during executive memory tasks (Osaka et al. 2007, Olson et al. 2009, Knops et al. 2009). In particular, the superior parietal lobule (SPL) which is known for its visuospatial tuning is shown to be critically involved in the memory manipulation (Koenigs et al. 2009). Moreover, in a series of fMRI studies, Yantis and his colleagues have shown that shifting of external attention between spatial locations and shifting attention in mnemonic domains show overlap across a fronto-parietal network Shomstein et al. 2006, Tamberrosenau et al. 2011,Chiu et al. 2009).

The result of these experiments have appeared in the following publications:

1. N. Noori, L. Itti, Spatial Registry Model: Towards a Grounded Account for Executive Attention, In: Proc. Conference on Cognitive Science (CogSci 2011), pp. 1-6, Jul 2011.
2. N. Noori, L. Itti, Eye-Movement Signatures of Abstract Mental Tasks, In: Proc. European Conference on Cognitive Science (EuroCogSci 2011), (B. Kokinov, A. Karmiloff-Smith, N. J. Nersessian Ed.), pp. 110:1-110:6, May 2011.
3. N. Noori, L. Itti, Visuospatial attention shifts during non-visual mental tasks, In: Proc. Vision Science Society Annual Meeting (VSS11), May 2011

Interference of executive memory tasks with Visuospatial Short Term Memory
Our Spatial Registry Hypothesis (SRH) predicts that visuospatial short-term memory might be implicated in those mental tasks that require manipulation of information. Thus their hypothesis implies that engaging with a mental executive task that requires memory manipulation not only impairs the visual perception but also will affect the spatial memory of previously perceived visual items. In particular SRH claims that an impact of a secondary non-visual mental task is independent of executive attentional load and selective to regions of space that is being used for the spatial registry.

Standard models of working memory explain the impact of mental tasks on the visual system commonly in terms of a bottleneck in executive attentional resources that are needed for both mental and visual tasks. However depends on one's assumption about the involvement of executive attentional resources in maintaining visuospatial short-term memory (VSSTM), an attentional-bottleneck account would predict either no impact or a uniform impact of a secondary executive memory tasks on VSSTM.

In these studies we investigated the possible involvement of the visuospatial short-term memory (VSSTM) in the manipulation of information in executive working memory tasks. We showed that performing a non-visual executive working memory task concurrent with a simple visuospatial memory task impairs the spatial short-term memory of visual targets selectively and independent of the load of the mental task. (Figures 3. and 4.)

We also verified this hypothesis by showing that the task-irrelevant spatial binding of items in a forward-or-backward recall task has a significant effect on the backward recall with demand for

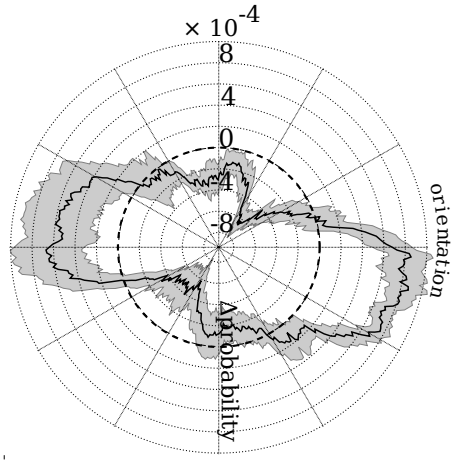
memory manipulation while it has no significant impact on the forward recall with no demand for memory manipulation.

In our first experiment we observed that subjects demonstrate a disadvantage in retaining visuospatial information along the vertical direction (the ignore condition of the first experiment ) (Figure 3.) . Therefore, if registering task-relevant items  with the visuospatial short-term memory play a functional role in the process of memory manipulation, one could imagine that using vertical direction for registering the items might result to a disadvantage in bookkeeping of memory items and hence might come into cost for the performance of the executive memory task.

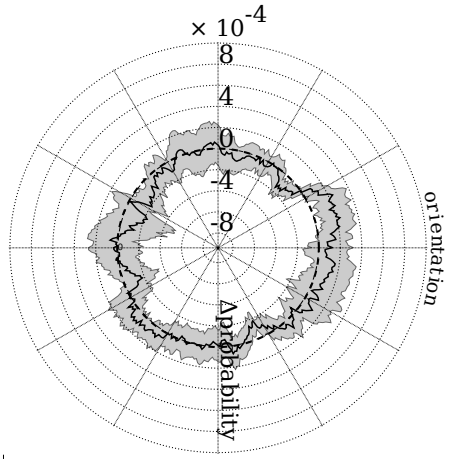
In our last experiment we investigated this prediction by priming subjects along two task-irrelevant orientations and monitoring their performance during an executive memory task .

We measured the performance of our subjects during a forward-or-backward recall task in which subjects first read some items from the screen without knowing whether they have to be recalled in a forward or a backward order. We measure the performance in both backward recall (executive memory task) and forward recall (active maintenance). The forward recall, presumably draws only on the phonological loop and thus the presentation method is not supposed to have any effect on the the performance. However backward recall requires memory manipulations and therefore is potentially sensitive to presumably task-irrelevant visual presentation method (Figure 5.).

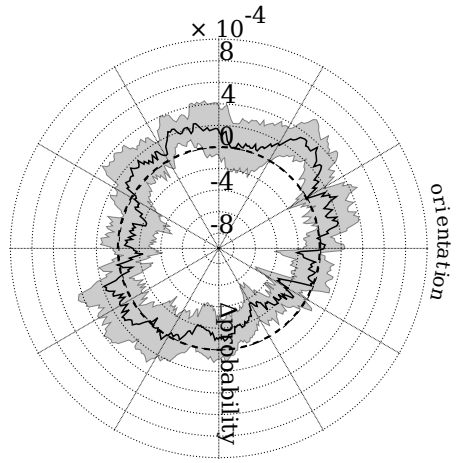
The result of these studies will appear in the following publication:
N. Noori, L. Itti, Selective Impact of Mental Abstract Tasks on Visuospatial Short-Term Memory, In: Proc. Vision Science Society Annual Meeting (VSS11), May 2012



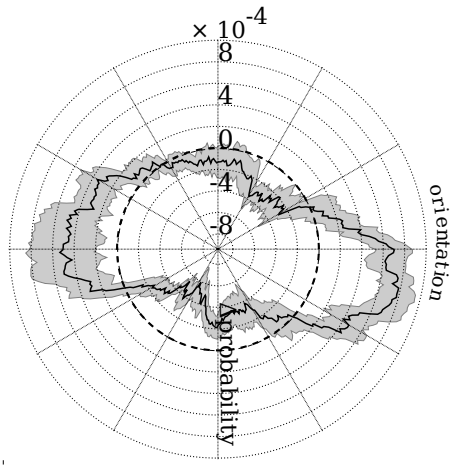

(a) Sorting : Horizontal - Vertical




(b) Maintaining : Horizontal - Vertical




(c) Sorting : Random - Horizontal




(d) Sorting : Random - Vertical

*Figure 1. Comparing the impact of active maintaining versus mental sorting on direction of gaze shifts. Each panel shows the subtraction of normalized distribution of directions of gaze shifts for a certain type of task and two different initial presentation method (a) when the mental task is sorting, the direction of initial presentation influences the direction dominant direction of gaze shifts. (b) the initial presentation has no impact of the directions of eye movement during active maintaining.(c) during mental sorting horizontal presentation and random presentation of items have the same effect.*

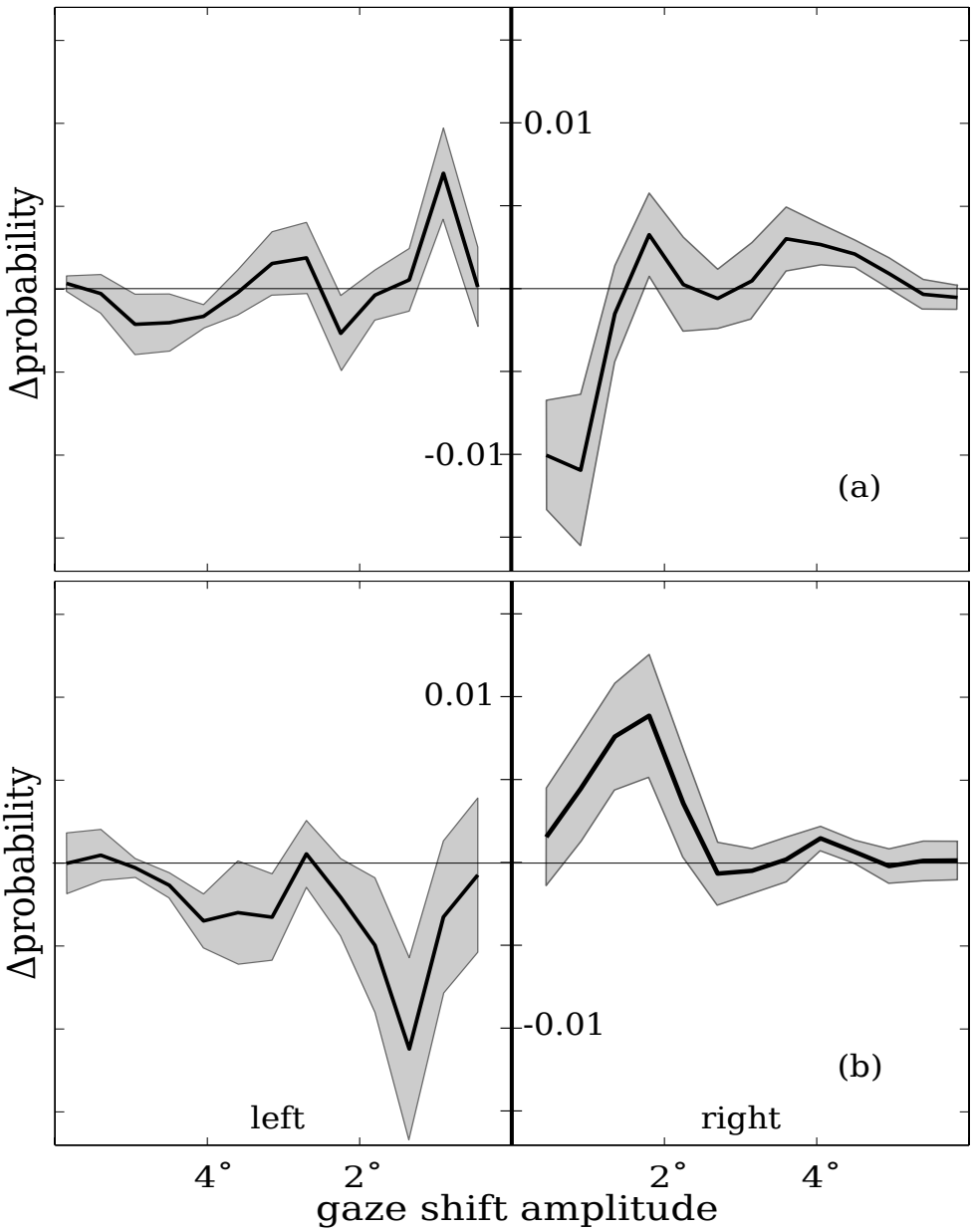



*Figure 2.Sequences of items which are symmetric induce symmetric gaze shifts during the sorting task. Each graph shows the difference between averages of normalized amplitude distributions of gazes for two symmetric sets of stimuli. On the top panel, the result for stimuli of type 1 – stimuli of type 2 (canonically represented by 34012 and 21043) is shown. The bottom panel shows the result for stimuli of type 3– stimuli of type 4 (canonically represented by 41230 and 03214).*

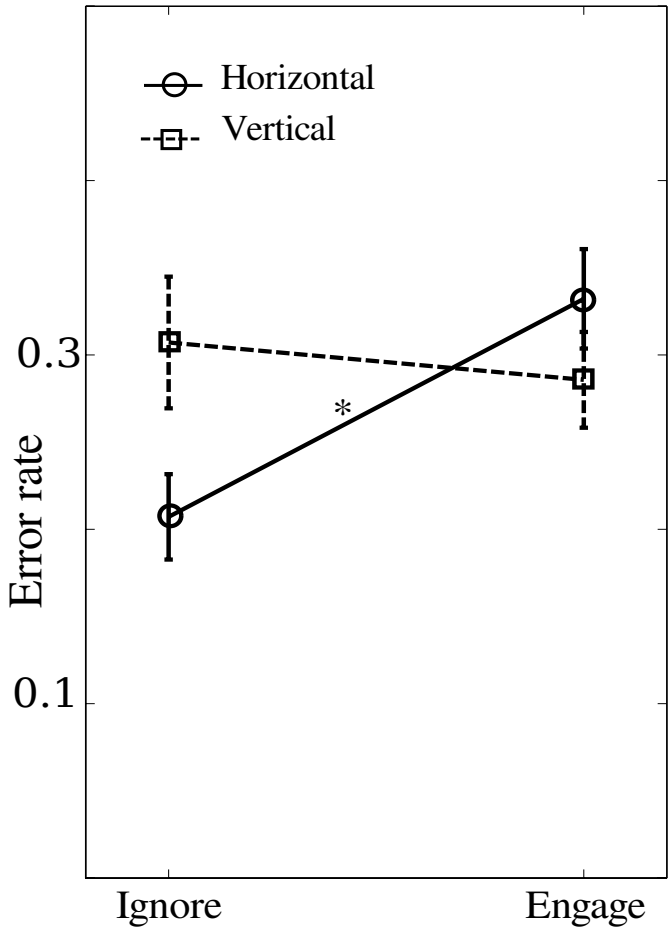



*Figure 3.Performing a mental sorting task impairs the visuospatial short-term memory along the horizontal direction while it has no effect on the visuospatial short-term memory along the vertical direction.*

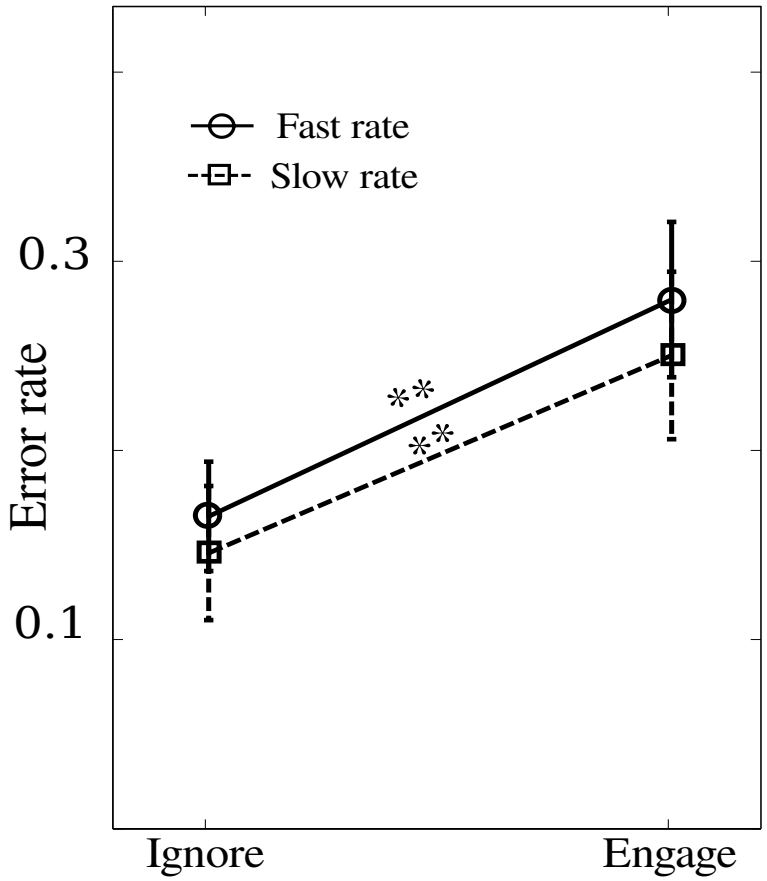



*Figure 4.Performing a double counting task impairs the visuospatial short-term memory. Howeverthis impact seems to be independent of the rate of counting.*

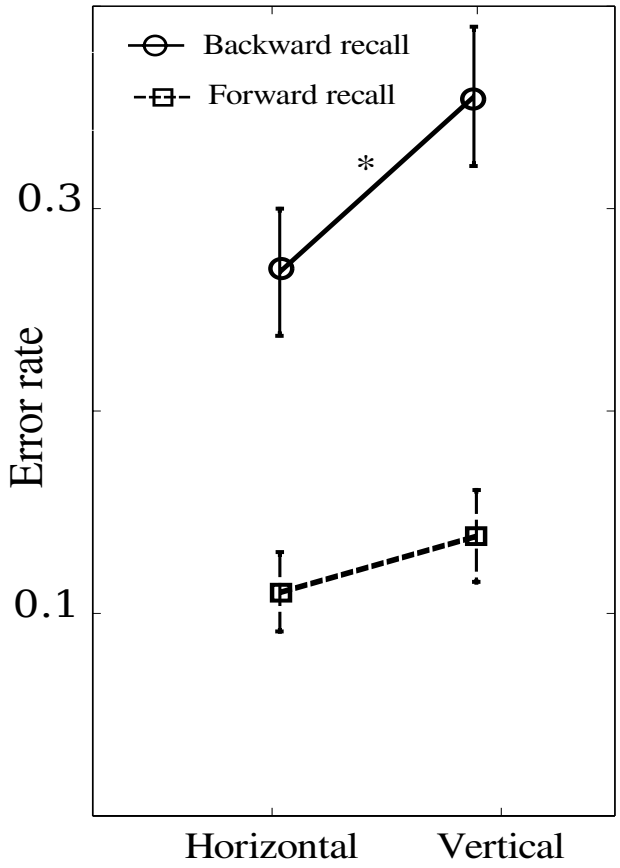



*Figure 5. The impact of the presentation direction on the errors of the forward and backward recall during a forward-or-backward recall task.*

# Bibliography


- Marr, D. Vision. In Vision, 1982.
- Ballard, D.H. Generalizing the hough transform to detect arbitrary shapes. Pages 714–725, 1987.
- Biederman, I. Recognition-by-components: A theory of human image. Psychological Review, 92:115–147, 1987.
- Chellappa, R. and Bagdazian, R. Fourier Coding of Image Boundaries. IEEE Trans. PAMI-6(1):102-105, 1984.
- Grigorescu, C., Petkov, N., and Westenberg, M.A.. Contour detection based on nonclassical receptive field inhibition. Image Processing, IEEE Transactions on, 12(7):729–739, 2003.
- Gustavsen, K. and Gallant, J.L.. Shape perception: Complex contour representation in visual area v4. Current Biology, 13:R234–R235, 2003.
- Huang, Chung-Lin and Huang, Dai-Hwa. A Content-based image retrieval system. Image and Vision Computing, 16:149-163, 1998.
- Hummel, J. E. and Biederman , I. Dynamic binding in a neural network for shape recognition.
- Isard, M. and Blake, A. Condensation - conditional density propagation for visual tracking. International Journal of Computer Vision, 29:5–28, 1998.
- Itti , L. and Baldi, P. F. Bayesian surprise attracts human attention. In Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005), pages 547–554, Cambridge, MA, 2006. MIT Press.
- Itti, L. and Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40(10-12):1489–1506, May 2000.
- Itti, L., C. Koch, and Niebur, E.. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254–1259, Nov 1998.
- Kanizsa, G (1955), "Margini quasi-percettivi in campi con stimolazione omogenea.", *Rivista di Psicologia* **49** (1): 7–30
- Kauppinen, Hannu. Seppanen, Tapio and Pietikainen, Matti. An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. IEEE Trans. PAMI-7(2):201- 207
- Lamme, V.A. and Roelfsema, P.R.. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci, 23(11):571–579, 2000.
- Lowe, D.G. Distinctive image features from scale-invariant keypoints. Intl. Journal of Computer Vision, 60(2):91–110, 2004.
- Lu, Guojun and Sajjanhar, Atul. Region-based shape representation and similarity measure suitable for content-base image retrieval. Multimedia Systems, 7:165-174, 1999.
- Maybeck, P.S. Stochastic models, estimation, and control. Academic Press, 1979.
- Medioni, G., Mordohai, P.: The Tensor Voting Framework. IMSC Press Multimedia Series. In: Emerging Topics in Computer Vision. Prentice Hall (2004) 191–252
- Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI), Acapulco, Mexico, 2003. IJCAI.
- Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. Fastslam: A factored solution to the

simultaneous localization and mapping problem. In AAAI, 2002.
- Murphy, K. Bayesian map learning in dynamic environments. In In Neural Info. Proc. Systems (NIPS, pages 1015–1021. MIT Press.
- Navalpakkam, V. and Itti, L. Search goal tunes visual features optimally. Neuron, 53(4):605–617, Feb 2007. Also see commentary / preview entitled “Paying Attention to Neurons with Discriminating Taste” by A. Pouget and D. Bavelier, Neuron 2007;53(4):473-475.
- Otterloo, Peter J. van. A contour- Oriented Approach to Shape Analysis. Prentice Hall International (UK) Ltd. C1991.
- Pasupathy, A. and Connor, C.. Responses to contour features in macaque area V4. Journal of Neurophysiology, 82 (5):2490–2502, 1999.
- Persoon, Eric and Fu, King-sun. Shape Discrimination Using Fourier Descriptors. IEEE Trans. On Systems, Man and Cybernetics, Vol.SMC- 7(3):170-179, 1977.
- Psychological review, 99:480–517, 1992.
- Qiu, F.T. and Heydt, R. von der Figure and ground in the visual cortex: v2 combines stereoscopic cues with gestalt rules. Neuron, 47(1):155–166, July 2005.
- Rauber, Thomas W. Two-Dimensional Shape Description. Technical Report: GR UNINOVA-RT-10-94, University Nova de Lisboa, Portugal, 1994.
- Sajjanhar, Atul. A Technique for Similarity Retrieval of Shapes. Master thesis, Monash University, Australia, 1997.
- Serre, T., Wolf, L., Bileschi, S.M., Riesenhuber, M., and Poggio, T.. Robust object recognition with cortex-like mechanisms. IMedioniEEE Trans. Pattern Anal. Mach. Intell., 29(3):411–426, 2007.
- Sobeland, I. G. Feldman. A 3x3 isotropic gradient operator for image processing. presented at a talk at the Stanford Artificial Project, 1968.
- Tu , Z., Chen, X., Yuille, A.L., and Zhu, S.C.. Image parsing: Unifying segmentation, detection, and recognition. In ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision, page 18, Washington, DC, USA, 2003. IEEE Computer Society.
- Tu, Z. and Zhu , S.C. Image segmentation by data-driven markov chain monte carlo. IEEE Trans. Pattern Anal. Mach. Intell., 24(5):657–673, 2002.
- Wolfe, J.M.. Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review, 1(2):202 – 238, 1994.
- Zahn, Charles T. and Roskies, Ralph Z.. Fourier Descriptors for Plane closed Curves. IEEE Trans. On Computer,c-21(3):269-281,1972.
- Zhu, S.C., Zhang, R., and Tu, Z. Integrating bottom-up/top-down for object recognition by data driven markov chain monte carlo. CVPR, pages 738–745, 2000.

OPEN ACCESS Freely available online

PLoS one

# Training Top-Down Attention Improves Performance on a Triple-Conjunction Search Task

**Farhan Baluch**[1], **Laurent Itti**[1,2]*

**1** Neuroscience Graduate Program, University of Southern California, Los Angeles, California, United States of America, **2** Department of Computer Science, University of Southern California, Los Angeles, California, United States of America

## Abstract

Training has been shown to improve perceptual performance on limited sets of stimuli. However, whether training can generally improve top-down biasing of visual search in a target-nonspecific manner remains unknown. We trained subjects over ten days on a visual search task, challenging them with a novel target (top-down goal) on every trial, while bottom-up uncertainty (distribution of distractors) remained constant. We analyzed the changes in saccade statistics and visual behavior over the course of training by recording eye movements as subjects performed the task. Subjects became experts at this task, with twofold increased performance, decreased fixation duration, and stronger tendency to guide gaze toward items with color and spatial frequency (but not necessarily orientation) that resembled the target, suggesting improved general top-down biasing of search.

**Citation:** Baluch F, Itti L (2010) Training Top-Down Attention Improves Performance on a Triple-Conjunction Search Task. PLoS ONE 5(2): e9127. doi:10.1371/journal.pone.0009127

**Editor:** Michael H. Herzog, Ecole Polytechnique Federale de Lausanne, Switzerland

**Received** December 16, 2009; **Accepted** January 15, 2010; **Published** February 18, 2010

**Copyright:** © 2010 Baluch, Itti. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** Supported by the National Geospatial-Intelligence Agency, the Defense Advanced Research Projects Agency, the National Science Foundation, and the Army Research Office. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: itti@usc.edu

## Introduction

Bottom-up, stimulus-driven processes as well as top-down, goal-driven processes exert influence on perception and therefore on the ability to perform visual tasks. Experts in a wide range of fields [1], from radiologists detecting tumors [2], image analysts screening baggage at the airport [3], pilots scanning their instrument panel [4], to chess grand masters [5] rely on their perceptual discrimination and selection abilities to make judgements often in life threatening situations. Tasks performed by these experts rely on both bottom-up and top-down processes to search for and direct attention towards features of the image that are crucial to enabling perceptual judgement with confidence. The central question in this study is whether, and to what extent, training and expertise improve, or otherwise modify, how rapid top-down goal-driven tuning of visual processing can enhance visual information for perceptual decisions, specially in feature rich enviornments.

Guidance of visual search for features in an image by top-down processes poses a constant demand on the visual and attentional systems to convert descriptions of desired target(s), which may change from moment to moment depending on behavioral goals, into appropriate guiding signals that can facilitate localization of a target. The quality of the guidance is determined by a number of factors including, i) the properties of the tuning functions of the sensory system [6], ii) the ability of the sensory system to eliminate noise [7], and iii) the discriminability of the target from distractors and background clutter (signal-to-noise ratio). On a short time scale, attention can enhance guidance through enhanced gain [8], enhanced spatial resolution [9], effective stimulus strength [10], or noise exclusion [7]. Analogous effects have been observed in perceptual learning studies over a longer time scale of up to a few days or longer.

Perceptual learning studies have shown that practice can improve performance in discrimination [11–14] and detection [15,16]. These studies have shown improvement in either a spatially or featurally specific manner and thus implicated early sensory cortex as the locus of plasticity and this has also been observed in electrophysiological studies [17,18]. Although most studies limit their training to either specific spatial locations or specific stimulus feature ranges, there has been some speculation about mechanisms of more general improvement in tasks. Some studies for example, have implicated the higher cortex [19–21] in learning. Plasticity effects have been observed in later visual areas, namely V4 and FEF (frontal eye fields), as a result of perceptual learning [22,23]. Learning in tasks such as visual search has also been shown to be less specific [24]. Sireteanu et al. [25] have shown non-specificity of perceptual learning effects specially in visual search tasks, and thus placed the locus of plasticity for learning a visual search task at a higher level than sensory cortices. One question which has remained outstanding, however, is whether training can improve the effectiveness of the dynamic top-down attention biasing process itself through what has been termed process-based learning [26], as opposed to exhibiting sharper visual discrimination abilities for a specific type of target or location (perceptual learning or automaticity through better memory retrieval [26]), or generally improving speed and/or performance on a task (task acquisition for search). This type of non-specific learning remains understudied and more specifically, the pairing of learning within a visual search task to observe the effects of training top-down attention remains relativity unexplored (although see [27]).

In this study we address the question of whether expertise can be gained in a triple-conjunction (color, spatial frequency, and orientation) search task when both the features and spatial location of the target are changed from trial to trial while maintaining a persistent level of bottom up uncertainty in the Shanon entropy sense. This imposes a novel and interesting new constraint on the type of learning that can occur, eliminating the cases of (perceptual) learning due to 'stimulus imprinting' [28] and focusing on what Goldstone [28] has termed 'attention weighting'. Specifically, this type of paradigm makes a demand on the observers to make fast trial-by-trial adjustments of top-down biasing weights in order to succeed in the search task. We also ask what difference, if any, training makes on the subjects' saccadic eye movements and the types of distractors that they look at. This is a departure from a typical learning paradigm where the stimulus set is often restricted in either space or feature set. We look for mechanisms of acquisition of general domain expertise when the observers are given a task that requires attention to the stimulus in order to achieve success. By analyzing eye movements we can ensure that effects beyond general task acquisition are captured. Changing the target on each trial puts the spotlight on mechanisms of attentional biasing efficacy rather than simple perceptual learning. We hypothesized that better biasing would lead to increased guidance towards items that are similar to the target as the biasing process would render items sharing features with the target more salient. Thus the number of items that were viewed need not necessarily be reduced but the quality of the set may improve. An alternate outcome would be that subjects view a smaller number of items which would suggest a trend toward automaticity or more pre-attentive guidance.

We show that learning occurs even when the target is changed in both features and spatial location on every trial. The improvement is marked by a decrease both in intersaccadic interval (ISI) and reaction time. The decrease in ISI suggests an improvement in discrimination and a stronger emphasis on the selection (detection) task. However, we did not observe a significant drop in saccade counts which suggests that the improvement in selection was limited to improving the 'quality' of the subset of items on the display that are scrutinized (the size of the subset remaining fairly consistent). We also find that subjects tend to exploit two of the three features of the stimuli, making saccades towards items that are similar to the target in color and spatial frequency but, interestingly, not necessarily in orientation.

In sum, our results provide evidence for a mechanism of expertise acquisition that is driven by production of better top-down biasing signals, the behavioral correlate of which is the increased similarity effect observed. This coupled with improved discrimination, likely driven by multiple exposures to the family of stimuli used in the task, define the enabling mechanisms that allow the transition from novice to expert.

## Methods

### Ethics Statement

Subjects gave written consent under a protocol approved by the Institutional Review Board of the University of Southern California, and were paid for participating in the study.

### Subjects

Human subjects recruited for this study were undergraduate and graduate students at University of Southern California. Subjects included four males and one female aged 21–26 years. All subjects had normal or corrected vision. Subjects gave written consent under a protocol approved by the Institutional Review Board of the University of Southern California, and were paid for participating in the study. Subjects were naive to the purpose of the experiment and had never seen any of the stimuli before.

### Stimuli

A set of colored Gabor patches were designed for this experiment, which provided the ability to vary features along three dimensions: color, spatial frequency, and orientation. The luminance profile of each Gabor patch is given by the following equation:

$$g(x,y,\theta,\phi) = e^{-\frac{x^2+y^2}{\sigma^2}} e^{(2\pi\phi i(x\cos\theta + y\sin\theta))} \quad (1)$$

where $\theta$ is the orientation of the patch, $\phi$ is the spatial frequency. Each patch subtended $4^\circ$ of visual angle. The phase of the sinusoid at each point was used to modulate the color of the pixels along the hue axis in the HSV color space, as shown in figure 1a. By sliding a window along the hue axis, the range of colors in the patch was changed, thus modifying the appearance of the patch. The window spanned from 0 to 360 and a hue shift essentially recentered the window around a given value. Each Gabor patch was then defined by its spatial frequency which ranged from 1.7 c/deg to 5.2 c/deg, orientation, which ranged from $25^\circ$ to $155^\circ$, and finally a color hue value that determined the shift of the hue window.

Search arrays were constructed from 32 Gabor patches embedded in 1/f noise in a 4×8 grid, with slight spatial jitter ($1^\circ$ along the x or y direction) applied to each patch. One of the Gabor patches was randomly chosen as the target for each search array.

### Paradigm

Subjects conducted 1,000 trials of visual search over the course of ten consecutive days. Each day consisted of a session of 100 trials with a break after 50 trials. Stimuli were presented on a large ($1920 \times 1080$ pixels) LCD monitor (Sony Bravia XBR-III) and subjects were seated in a comfortable chair with their head stabilized by a chin rest. The viewing distance was 97.8 cm, corresponding to a field of view of $54.8^\circ \times 32.7^\circ$. A typical trial, as illustrated in figure 1b, began with a fixation cross at the center of the display followed by a 2 second target preview, presented at the center with a gray background. The gray value of this background was equal to the mean gray of the $1/f$ noise of the corresponding search array display. Subjects were instructed to find the target as fast and accurately as possible and had a maximum of ten seconds to find the target. Their eye movements were recorded as they searched for the target (see below for eye-tracking methods). Upon locating the target, subjects pressed a response button, at which point the search array disappeared. A display consisting of numbers that corresponded to the Gabor patch locations was then displayed for 200ms. Subjects had to read and key-in the number at the location of the target using a keyboard. The font size was sufficiently small that one could not read the numbers corresponding to one Gabor patch while fixating at the location of any other Gabor patch. The goal of this 'no cheat' procedure was to ensure that subjects reported correctly the patch which they thought was the target(for more details on this procedure see [29]). After subjects provided input, they were given feedback as a 'correct' or 'incorrect' response, as well as the current level of performance (% correct responses so far). Each session lasted approximately 45 minutes.

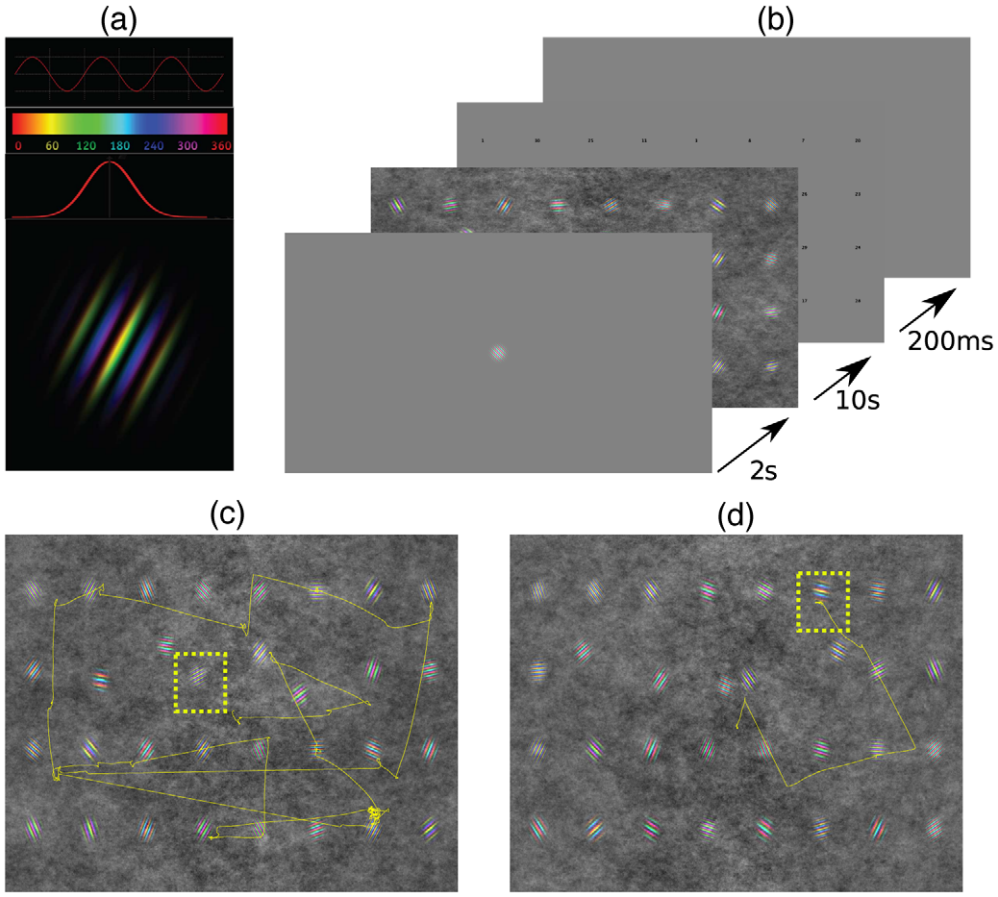



**Figure 1. Stimulus and Paradigm.** (a) Color Gabor patches constructed by first applying a gaussian envelope over a sinusoid as shown. At each point the phase of the sinusoid was used to modulate a hue axis in the HSV color space. (b) A trial started with a two-second target preview followed by a display of the search array for a maximum of ten seconds. If subjects found the target before the 10 seconds elapsed they hit a key to move to the next display. The next display showed numbers corresponding to Gabor patch locations in the search display. The numbers were displayed for only 200ms to ensure that subjects fixate the target in order to report the correct number. Subjects were then aske d to report the number at the target location. (c) A typical eye trace overlayed on a search array, showing an early trial. (d) A typical eye trace overlayed on a search array, showing a late trial.
doi:10.1371/journal.pone.0009127.g001

## Stimulus Presentation and Eye-Tracking Procedures

The subjects' eye movements were recorded as they searched for the target in the search array. Eye movements were recorded at a sampling frequency of 240 Hz, using an infrared-video-based eye-tracker (ISCAN RK-464) and the pupil and corneal reflection from the right eye were used to determine the gaze position with an accuracy of $\leq 1^{\circ}$. Calibration was performed using an online system that presented subjects with a central fixation point followed by a point at one of nine locations on a $3\times3$ grid. Subjects had to saccade from the central fixation point to one of the nine locations and maintain stable fixation (x and y position variance $<5$ pixels) for 300ms (75 samples). Once stable fixation was established the next location was presented. This process was repeated until stable fixations at all nine points were found. The eye positions obtained were then used to perform an affine transform and the transformed eye positions were displayed on the screen for the experimenter to confirm that an accurate calibration session had been conducted. During offline analysis a further thin-plate-spline interpolation [30] was performed to obtain accurate transformation from eye-tracker coordinates to screen coordinates. A recalibration session was performed every 20 trials to correct for possible head movements. Once transformed, the eye-traces could be overlaid on the images for further analysis as shown in figure 1(d).

## Data Analysis

The subjects' eye movements were calibrated as described above and an algorithm was used to parse the eye movements into saccades using a combination of filtered instantenous velocity measurements and a simple windowed Principal Components Analysis (PCA). Eye movement segments with a minimum velocity $30^{\circ}/s$ and a minimum amplitude of $2^{\circ}/s$ were classified as saccades. Blinks were identified by a pupil diameter reading of zero and trials with either blinks or loss of tracking for more than 10% of the trial were removed from further analysis. Unfortunately, on day two, one of the subjects' eye movements were lost due to machine failure; however, he completed all trials and continued to participate in the study. This loss not withstanding, we retained 97% of the 4,900 available trials, obtaining a total of 76,287 saccades for analysis.

We performed analysis on changes over time in the subjects' eye movements by constructing feature similarity maps and correlating these with binary saccade maps. The feature similarity maps were constructed as follows. We first discretized the feature space by dividing each dimension into ten bins (several numbers were tried for this and numbers between 10–25 bins gave similar results). Each Gabor patch was then defined as a triplet of bin values $G_i=\{h_i,f_i,o_i\}$ where $h_i,f_i,o_i$ are the bins of hue, frequency, and orientation respectively of Gabor patch $i$. A feature similarity map

for each trial consists of 32 cells arranged in a 4×8 grid each corresponding to one of the color Gabor patch in the search array for that trial. Similarity maps for each feature were constructed individually. A feature similarity map for hue, for example would contain in each cell $i$ the difference between the hue bin value $h_i$ of the Gabor patch and the hue bin $h_t$ of the target Gabor patch $t$ for the trial. In order to maintain an intuitive sense of the similarity measure (high values for high similarity) we computed similarity between the target patch $t$ and a Gabor patch $i$ for each feature $f$ as $s^f_{it} = -|f_i - f_t| + granularity$ (where granularity was set to ten since we divided the feature space into ten bins). Large values in cells therefore mean that the particular distractor was very similar to the target and vice versa.

As described before we drew the features of the distractors in each display from a uniform distribution and therefore by design the bottom up uncertainty in each display averaged across sessions should remain constant. In order to ensure that this was the case we computed the Shannon entropy in each feature similarity map. This enabled us to quantify the amount of uncertainty in our arrays. We then computed the average entropy per session and ran a regression to look for any trends over time. As expected we found no significant trends (color $r^2 = 0.03, p = 0.63$; frequency $r^2 = 0.12, p = 0.33$, orientation $r^2 = 0.22 p = 0.17$).

To construct binary saccade maps we first assigned saccade end points to Gabor patches if the distance from the end point to the center of the Gabor patch was smaller than 3.5°. These assignments allowed us to fill a 4×8 grid of cells corresponding to the 4×8 grid of Gabor patches, with 1 for a saccade end point landing on the Gabor patch and a 0 for no saccade towards the patch. In this manner binary saccade maps were constructed and later correlated with the feature similarity maps. When a particular patch was fixated several times we still placed a one in the map in order to retain the binary nature of the saccade maps.

## Results

### Performance

Measuring performance as the percentage of correct trials for each 100-trial session, we found that subjects showed improved performance over the course of the trials (figure 2). The mean percentage performance of the group was computed by taking an average of the percentage correct responses by each of the five subjects for each session. A one-way ANOVA showed an effect of session on mean performance ($F(9,40) = 6.88$ $p < 0.01$). The change in performance measured by the slope (indicative of learning rate) of the logistic fit on the data halfs at day five and later levels off, hovering around 70% to 80% correct as shown in figure 2.

This indicates that the subjects improved on the task and answered correctly a greater percentage of time after conducting several hundreds of trials of the task, despite the fact that the features and spatial location of the target was changed on every trial. Pooling together the reaction times for each subject and averaging across the sessions revealed an effect of session on the mean reaction time (figure 3a) for our pool of subjects (one-way ANOVA $F(9,4990) = 50.71$ $p < 0.01$). A similar but weaker effect in number of saccades was observed (one-way ANOVA $F(9,4766) = 12.62$ $p < 0.05$) as shown in figure 3c. To ensure that the performance improvements observed were not due to a speed-accuracy tradeoff, we normalized performance by the mean number of saccades and mean reaction time separately. Mean performance normalized by the mean number of saccades gave us a measure of subjects' per-saccade search efficiency. Plotting this as a function of sessions (figure 3d), we find an increased per

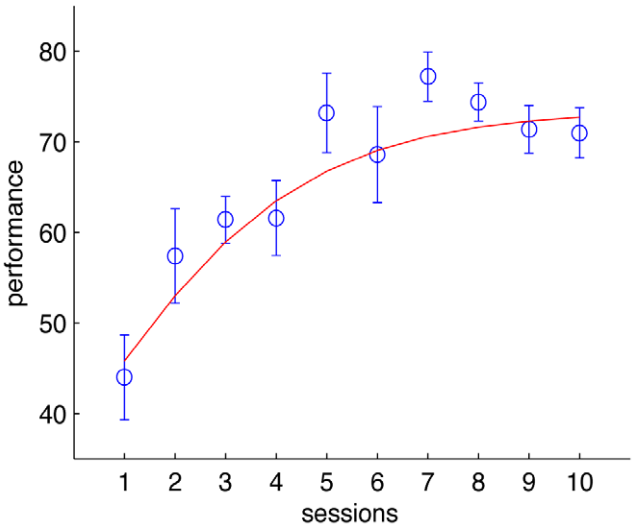



**Figure 2. Performance results.** Mean percentage correct performance obtained by taking a mean across subjects for each of the 10 sessions. Error bars are SEM across subjects. Smooth curve is a fit to a logistic function ($r^2 = 0.62, p < 0.05$).
doi:10.1371/journal.pone.0009127.g002

saccade efficiency (one way ANOVA $F(9,40) = 2.43$ $p < 0.05$). Similarly, plotting mean performance (figure 3b) per session normalized by the mean reaction times we find an upward trend of search performance per unit time spent searching (one-way ANOVA $F(9,40) = 3.71$ $p < 0.01$). These results show a clear improvement of all subjects on the task with training. To confirm that learning was not just a result of improvement in reporting the numbers in the brief display, we examined the accuracy of reporting the number at the position last fixated. We found that the number at the position of last fixation matched the reported number 82.6% of the time on incorrect trials and 92.8% on correct trials. Further pooling the trials together and computing an average over each session, normalized by the number of incorrect trials, we find no effect of session on report accuracy (one-way ANOVA $F(9,40) = 0.77$, $p = 0.65$). Thus, we can rule out that performance improvements might have been due to an improved ability to read and report the numbers.

### Differences in Basic Eye Movement Statistics

The eye movements of all the subjects were grouped by session, and statistics were then computed on this data. We first analyzed the main sequence, which plots peak velocity against saccadic amplitude. The main sequences for session one and session five are shown in figure 4a. To determine whether there was a difference between the two sequences we first fitted a linear function to the main sequence of session one and then used this model to predict saccade amplitudes using the peak velocity data from session five saccades. We then ran a two-sample t-test between predicted saccade amplitudes and real saccade amplitudes for session five and found no significant difference ($p = 0.50$). The analysis of the main sequences therefore revealed no effect of training on these saccade statistics, and the subjects' eye movements were similar in this regard. Similarly, no significant trend was found in saccadic amplitude or velocity individually (data not shown). However, when we analyzed the ISI we found a significant drop from early sessions in training to late sessions, as illustrated in figure 4b. Specifically, a one-way ANOVA showed a strong effect ($F(9,73481) = 43.95$, $p < 0.05$) of session on intersaccadic interval.

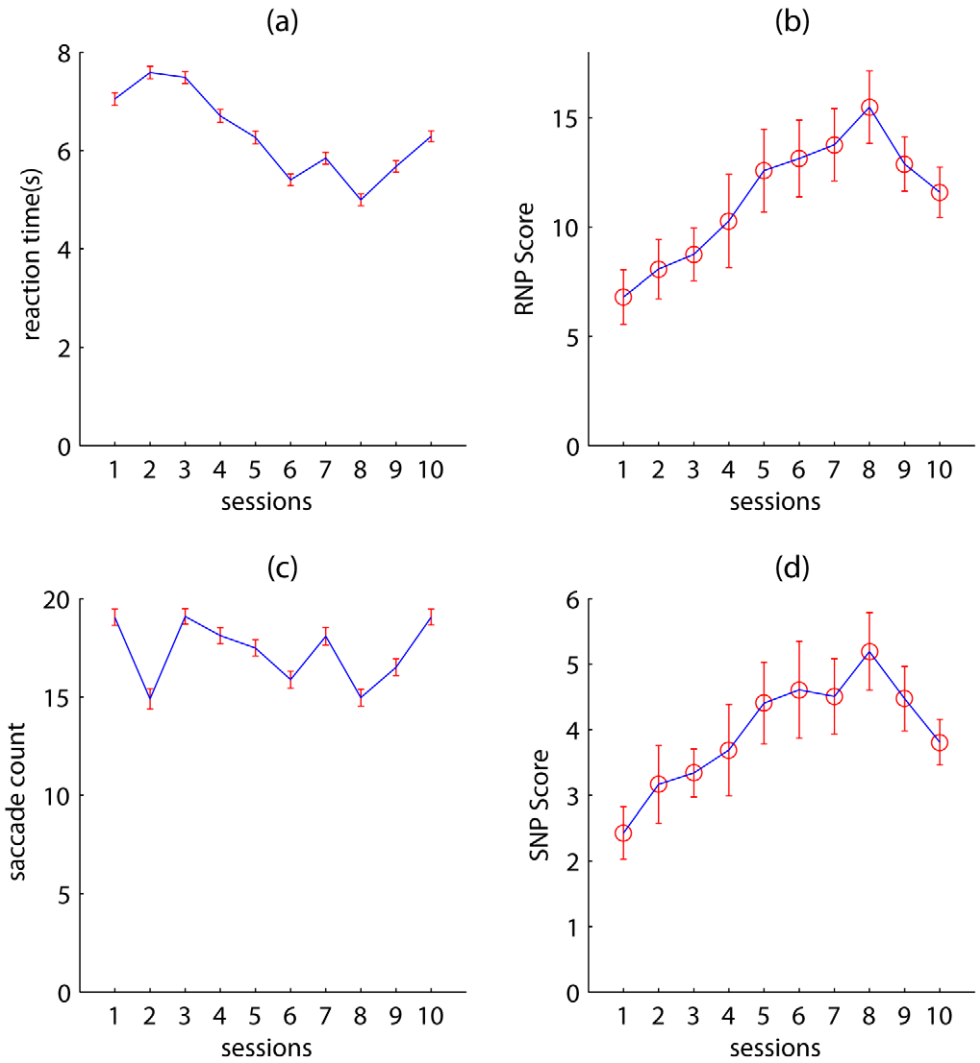



**Figure 3. Reaction time and saccade count data.** (a) Reaction time plotted as a function of session computed by pooling together all trials by all subjects for each session and taking the mean. Errorbars are SEM. (b) Reaction time Normalized Performance (RNP) score computed by normalizing mean performance by mean reaction time per session. Error bars are SEM taken across sessions. (c) Saccade counts plotted as a function of session, computed by pooling together data from all subjects per session and taking a mean. Errorbars are SEM. (d) Saccade count Normalized Performance (SNP) score computed by normalizing mean performance by mean saccade count per session. Errorbars are SEM.
doi:10.1371/journal.pone.0009127.g003

These results demonstrate a change in saccadic strategy on the part of the observes, a change marked by increased efficiency in examining the Gabor patches and greater speed in rejecting non-target Gabor patches. As expected a fall in ISI resulted in a drop in reaction time (RT). However, we found that RT was more strongly dependent on the number of saccades made rather than on ISI. We found a significant dependence ($r^2 = 0.69, p < 0.05$) of RT on the number of saccades made (figure 4c). A weaker dependence (figure 4d) of RT on ISI was found ($r^2 = 0.18, p < 0.05$). The data shown in the figures is for trials where reaction time was $< 10s$; the results for the full dataset were similar (RT vs saccade count $r^2 = 0.57, p < 0.05$ and RT vs ISI $r^2 = 0.22, p < 0.05$). Therefore number of saccades appeared to be more important in determining RT than ISI.

## Individual Feature Similarity Map and Saccade Map Correlations

Having constructed feature similarity maps and binary saccade maps, a correlation value between the binary saccade map and each of the feature correlations maps were computed for each trial. Correlation values for each session were computed by pooling together trials of all subjects within a session and then computing the mean. Figure 5 shows that, i) feature similarity maps and binary saccade maps are correlated, and ii) hue and frequency similarity maps become increasingly correlated as the sessions progress, however, no such trend can be observed for orientation. The positive trend indicates correlations between non-zero values in the binary saccade map with high values in the feature similarity maps. This demonstrates a higher likelihood of subjects making saccades towards items that are similar to the target.

The significant increase in correlation of the hue map from session one to session five (paired t-test $p < 0.05$) demonstrates that subjects increasingly looked at items that were closer in hue to the target. There was also a significant increase in frequency correlation from session one to session five (paired t-test $p < 0.01$), once again demonstrating a tendency to saccade towards items with frequency more similar to the target. This

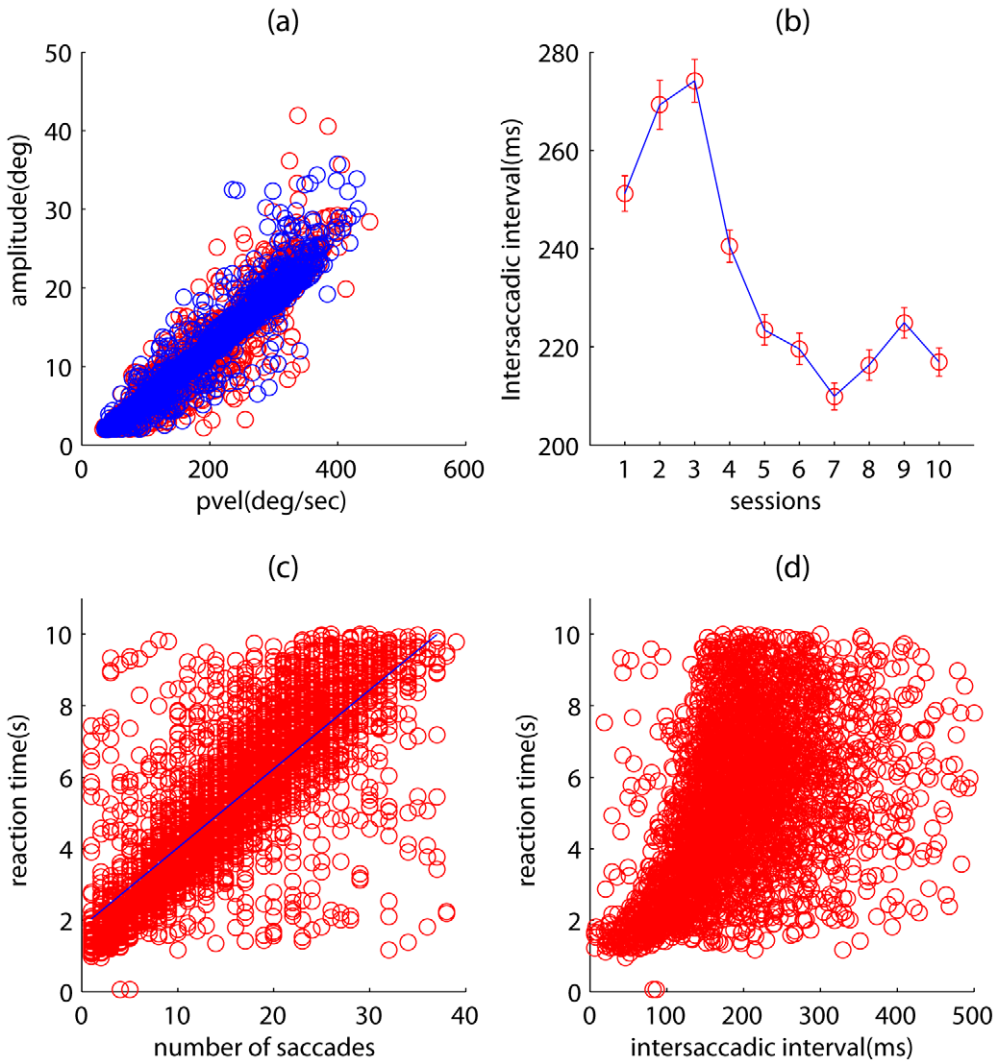



**Figure 4. Saccade statistics.** (a) Main sequence, plotting saccade amplitudes against peak velocity for the first session (red) and fifth session (blue). Overlap shows no difference in main sequence. (b) Intersaccadic interval reduces with session data. Points were computed by pooling saccades for each session for all subjects and taking a mean. Error bars are SEM. (c) Reaction time as a function of number of saccades. Regression line shows significant correlation ($r^2 = 0.58, p < 0.05$). (d) Reaction time as a function of intersaccadic interval. Regression shows weak correlation ($r^2 = 0.22, p < 0.05$).
doi:10.1371/journal.pone.0009127.g004

was not the case for orientation, where we found a non-significant ($p = 0.36$) difference between session one and session five.

We further quantified this result by running a multiple logistic regression on the data, examining the combined effect of feature distances on the probability of making a saccade towards the target in a given session. Coefficients obtained from this regression were then plotted as a function of session and fitted to a logistic function $y = \frac{L}{1 + ce^{-ax}}$ (figures 6a, b, and c), where $L$ is the upper limit of the curve, and $a$ determines the slope of the curve, while $c$ determines shift of the inflection point of the function. $L$ is evaluated by computing an average of the coefficient values for sessions seven through ten. The coefficients' trends plateau at seven coinciding with a plateau in performance thus we use the mean to compute $L$. We then linearized the function to run a linear regression that provided a method for computing the parameters $c$ and $a$. The regressions yielded significant trends for hue ($r^2 = 0.50, p < 0.05$), and frequency ($r^2 = 0.49, p < 0.05$) coefficients but not for orientation ($r^2 = 0.18, p = 0.2216$).

These results demonstrate a tendency of subjects to exploit hue and frequency as the primary features while giving lowest priority to orientation. This effect has also been observed in previous studies [31–33] that found a hierarchy of feature efficacy in biasing saccades towards targets, with color being the dominant feature followed by size and orientation.

## Feature Combination Rules

We also investigated the question of what combinations of features might be learned. Several feature combination rules were tested by combining the similarity maps using different computations. Figure 7 plots the correlation values across the sessions for maps constructed using various methods of combining the individual feature maps. A linear combination rule for individual features is most widely used [34,35] where individual features are combined through a linear operation to form a final saliency map that guides

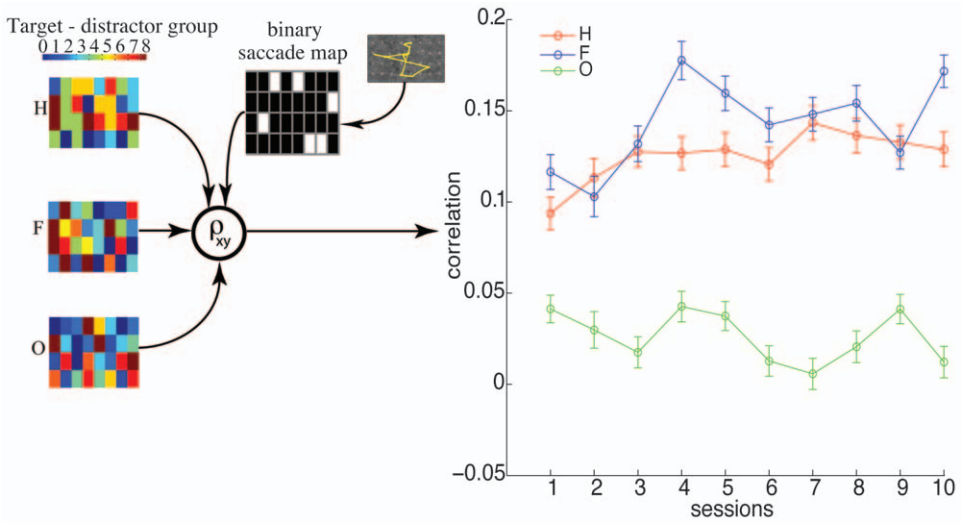



**Figure 5. Single feature correlations.** Feature similarity maps are shown on the left with hot colors showing high similarity. These similarity maps are correlated with saccade maps to yield a correlation value $\rho_{xy}$. The plot shows mean correlations per session for each feature. Error bars are SEM.
doi:10.1371/journal.pone.0009127.g005

attention. Top-down attention has been hypothesized to modulate the contribution from each map in an optimal manner [36] by adjusting biasing weights [37,38]. Correlation between binary eye movements maps and feature similarity maps constructed by combining linearly the hue, frequency, and orientation similarity maps (appropriately weighted) should therefore be high.

We constructed similarity maps by linearly summing the individual feature similarity maps for all combinations of the

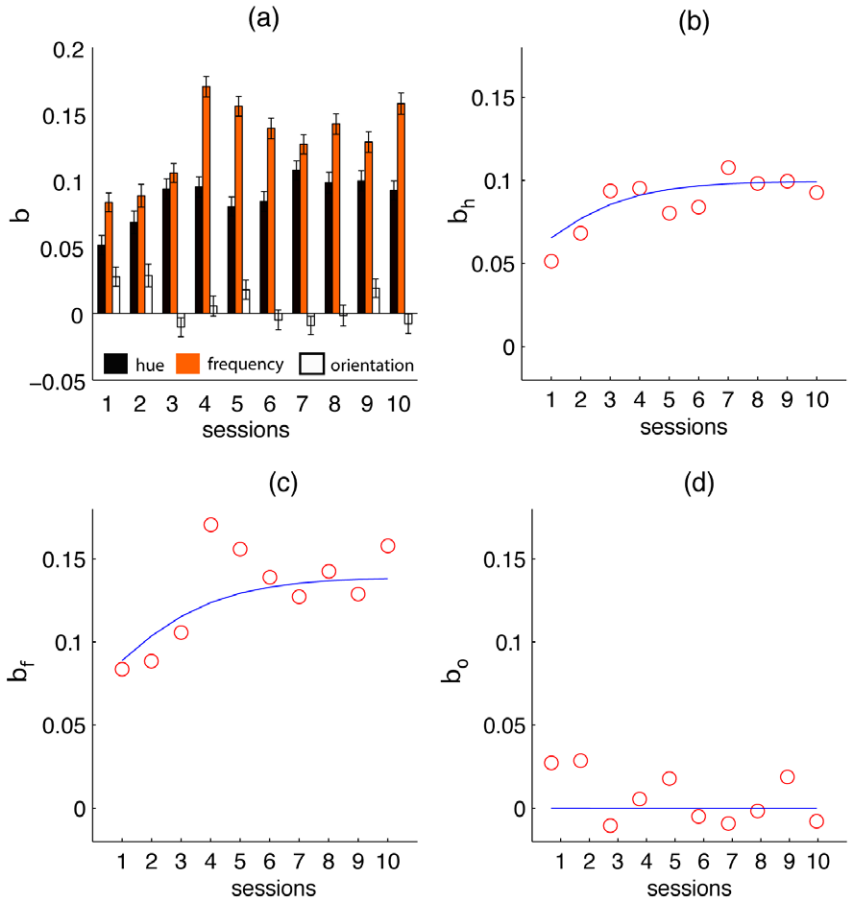



**Figure 6. Multiple Logistic regression results.** (a) Coeffecient values for each feature plotted as a function of session. (b) Regression line fitted to the coefficient values for hue ($r^2 = 0.50, p < 0.05$), (c) frequency ($r^2 = 0.49, p < 0.05$) and, (d) orientation ($r^2 = 0.18, p < 0.2216$).
doi:10.1371/journal.pone.0009127.g006

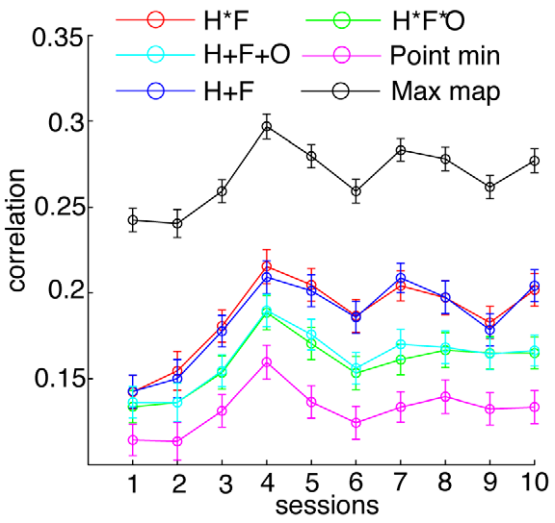



**Figure 7. Feature combination correlations.** Plots showing correlations of feature similarity maps combined using various methods, as a function of sessions. The black curve (Max map) represents an upper bound computed by taking the most correlated feature map on each trial and computing averages across all trials for each session. The correlation values for this upper bound can be used to compare mean correlation values for all other combination rules H*F (red), H*F*O (green), H+F (blue), H+F+O (cyan) and, point wise minimum rule (magenta).
doi:10.1371/journal.pone.0009127.g007

three features, and found that the map formed from a linear combination of the hue and frequency maps (H+F), was most strongly correlated with eye movements.

To obtain an upper bound of correlation against which each rule in figure 7 could be compared, we created a maximum map (labeled "MaxMap" in the figure). The correlation values for this map were computed by taking the feature similarity map on each trial that had the strongest correlation with the saccade maps and storing this correlation value. The mean across trials was then computed from this trial-wise maximum, thus yielding an upper bound. We found that the map formed from the linear combination of hue and frequency (H+F map) was the closest to the upper bound. A significant effect of session on correlation values for this map was also observed (one-way ANOVA $F(9,4666)=6.61 p<0.05$). This suggests that subjects attended to the hue and frequency features and improved on the task by appropriately tuning top-down signals in the hue and frequency dimensions.

We also explored a multiplicative combination rule whereby we combined the maps in a point-wise multiplicative manner. Thus if a feature at a particular location is poorly matched to the target's feature it will eliminate the chance for all other features to select this location as a potential target. This predicts a sparse saliency map, and has the elements of an AND operation on the multiple feature maps. However, if we look at the correlation values for the multiplicative map H*F*O they are not as strongly correlated as the H+F map. Despite the weak correlation we do find a trend in the correlation values for the H*F map (one-way ANOVA $F(9,4666)=5.61 p<0.05$). These results demonstrate a general improvement in the subjects' tuning to the features of the target upon preview and also suggests that while the multiplicative rule makes for a computationally useful guidance strategy, a linear rule may be a more biologically plausible operation.

We then constructed a point-wise minimum map which would have the highest signal-to-noise ratio. The map was constructed by placing in each cell the value of the least similar item. In this manner the map contains low values in all locations except at the target cell location where the three feature maps would contain equal values. This strategy would call on a hypothetical observer to adopt the counter-intuitive strategy of searching for features that are most dissimilar to the target, thus highlighting a single location (target location) where no dissimilarities are found. However, it is difficult to conceive of a neural strategy that would enable such a mechanism since it would require pre-computation of all three feature maps, extraction of the most discriminative feature for each item, followed by construction of the final guidance map.

## Discussion

The triple conjunction search task learned by subjects in this study consisted of displays that remained consistent in the number of items and bottom-up uncertainty, however, the target changed both its location and features on each trial. Learning still took place under these conditions and the combined behavioral, occulomotor, and perceptual signatures of the improvement point towards effects beyond task acquisition. Behaviorally we saw an improvement in performance with subjects reporting the correct target on average 44% of the time at the beginning of the task to an average of 71% after developing expertise in this feature-rich environment. The occulomotor correlate of learning was evident from the changes in saccadic behavior, namely in the shorter ISI with training. Differences in basic saccade statistics in conjunction with visual search as well as learning have not been studied extensively. Phillips et al. [39] argue that gains in visual search performance are a result of an expansion in the 'perceptual span' and forward saccade amplitude, with a small effect of fixation duration which is equivalent to the ISI in our case. The improvement obtained in our case suggests both that there was an increase in perceptual span, as well as reduced dwell time for extracting information from each fixation.

Hooge & Erkelens [40] conducted experiments to specify the role of fixation duration in visual search tasks. The most salient feature of their study was the reconciliation of contradictory findings of [41] who found significant guidance of saccades towards items that were similar in color to the target, and Zelinsky [42] who did not find such guidance. Hooge & Erkelens [40] provide a means to make a leap from occulomotor dynamics to visual search performance using fixation duration as the vehicle for understanding the difference. They suggest that tasks involving difficult discriminations but easy peripheral selections tend to invoke longer fixation durations, while tasks involving easy discrimination but difficult peripheral selection (due to either an abundance or similarity of distractors around a target) tend to have shorter fixation durations but evoke a greater number of saccades. Our task is a difficult conjunction search where distractors share features with the target, this makes it a 'hard-discrimination, hard-selection' task. Therefore, initially we obtain high ISI's (in fact ISI goes up from session one to session two) which perhaps suggests that our subjects' occulomotor strategy focused on the foveal discrimination early in the task. High saccade count and reaction times suggest that the selection task was not easy either. However, with training we obtain much lower ISIs which implies that subjects improved on the discrimination task and could now concentrate resources on the selection task. Further, we find that the mean number of saccades stays fairly constant with subjects scanning over half the number of items on average. Thus, there is no significant change in the number of selections made during the search process, however, the 'quality' of the selections improves, i.e. the distractors chosen as potential targets are closer in their

features to the target. The quicker ISIs may point toward an increased 'perceptual span' [43] or 'visual lobe' [44] that enables examination of a greater number of items in each saccade, however, additional experiments would be required to confirm this claim.

The occulomotor correlate of learning (i.e. improved discrimination by moving from discriminative search to selective search) then makes the prediction that subjects would have a higher tendency to make saccades towards patches that are similar to the target as they transition from discriminative search to selective search. Indeed this is what we found when we correlated saccade maps with feature similarity maps. By running a multiple logistic regression we found that whether a patch was selected for fixation could be predicted by the similarity of its features to the target and level of training of the subjects. These results on the similarity effect [45] serve as corroboration of several previous studies including [31] who found that monkeys make fixations to items that are similar in color but not orientation. Findlay & Gilchrist [45] also found a proximity effect, i.e., a tendency of saccades to fall near the target in space. Motter & Belky [31] also investigated this selection for color as a guiding feature over orientation. They conclude from their 1998 study, as well as electrophysiological studies in V4 [46,47], that V4 neurons coded more strongly for stimuli in their receptive field that matched the top-down goal rather than the absolute color of the stimuli. This suggests that a color feature map would be the tool of choice for top-down attention in the guidance of saccades. Our study also demonstrates a preference for spatial frequency over orientation. Several other studies [32,33] have found a similar preference for color as a guiding feature, and Wolfe & Horowitz [48] have placed color on top of the list of features that guide attention. We hypothesize that spatial frequency could be considered a 'surface property' much like texture and color that have desirous qualities for the guidance of attention. However, the current experiment does not address this feature-selective guidance and it would require further experiments to verify why orientation is a weaker cue for top-down attention in the presence of other features.

In this study the top-down goal changed on each trial and despite this we saw an increased similarity effect which suggests that activity of neurons in the visual cortex (e.g. V4 neurons) can be biased in a highly dynamic and rapid manner from one trial to the next. Therefore departing from typical perceptual learning studies we show evidence for learning that involves top-down processes. Herzog & Fahle [49] put forward a recurrent neural network model of perceptual learning that empahsizes the role of plasticity in the top-down connections as an enabling process for perceptual learning. They show that even in a task like vernier discrimination, where learning is both specific to stimulus features and spatial location, a model that incorporates top-down influences has more explanatory power than pure bottom-up models of improvement. Specifically they show that in a model where top-down connections gate flow of bottom-up inputs to decision units, learning acts upon the weights of the top-down connections rather than tuning properties of the bottom-up (sensory) inputs. The current study can also be placed in this context, situating the locus of plasticity in the top-down process rather than the bottom-up sensory process. However, in addition to this the increase in the similarity effect that we find, suggests that the ability to quickly switch the top-down signal also improved. It is certainly the case that there is a task-based effect and we cannot ascertain the exact amount of contribution which exclusive improvement in top-down biasing made toward progress in the task. However, it is clear from our analysis of correlation between feature similarity maps and binary saccade maps that there is enhanced guidance through better top-down biasing. We find that training enhances the similarity effect and a possible mechanism for this is improved top-down biasing. This enhances the right neurons which in turn guides attention to patches that are increasingly similar to the target.

Conjunction searches define targets using a combination of features, and binding of these features according to feature integration theory [34] requires attention. We examined the correlations of binary saccade maps and different combinations of feature similarity maps and found that a linear combination of the features hue and frequency was most highly correlated with saccade maps. We tried a multiplicative rule which provides the sparsest final similarity since it penalizes differences in a single feature while greatly boosting locations with a single matched feature. A similarity map constructed from a multiplication of hue and frequency was closely matched in terms of correlation with eye movements to the linear H+F map however, the H*F*O map was poorly corrleated with eye movements. A multiplicative rule however, does not account for the serial search times for conjunction searches since a precomputation of this multiplicative combination of features would put a hot-spot in a salience map at the location where all features match the target with high SNR. Overall this exploration points towards a linear combination rule that may be at play. That said, our discussion of the similarity effect also suggests a pre-attentive guidance of saccades towards potential targets. And if guidance is pre-attentive and feature combination requires attention, the prediction would be that conducting a conjunctive search is a serial process with respect to spatial attention and feature-based attention, and thus inefficient.

## Author Contributions

Conceived and designed the experiments: FB LI. Performed the experiments: FB. Analyzed the data: FB. Wrote the paper: FB LI.

## References


1. Kellman P, Garrigan P (2009) Perceptual learning and human expertise. Physics of Life Reviews 6: 53–84.
2. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, et al. (1988) Expertise in a complex skill: Diagnosing x-ray pictures. The nature of expertise. pp 311–342.
3. McCarley J, Kramer A, Wickens C, Vidoni E, Boot W (2004) Visual skills in airport-security screening. Psychological Science 15: 302–306.
4. Bellenkes AH, Wickens CD, Kramer AF (1997) Visual scanning and pilot expertise: the role of attentional flexibility and mental model development. Aviat Space Environ Med 68: 569–79.
5. Ferrari V, Didierjean A, Marmèche E (2008) Effect of expertise acquisition on strategic perception: the example of chess. Q J Exp Psychol (Colchester) 61: 1265–80.
6. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18: 193–222.
7. Dosher BA, Lu ZL (2000) Noise exclusion in spatial attention. Psychological Science 11: 139–146.
8. Treue S, Maunsell JH (1996) Attentional modulation of visual motion processing in cortical areas mt and mst. Nature 382: 539–541.
9. Yeshurun Y, Carrasco M (1998) Attention improves or impairs visual performance by enhancing spatial resolution. Nature 396: 72–75.
10. Reynolds JH, Pasternak T, Desimone R (2000) Attention increases sensitivity of v4 neurons. Neuron 26: 703–714.
11. McKee S, Westheimer G (1978) Improvement in vernier acuity with practice. Perception & Psychophysics 24: 258–62.
12. Vogels R, Orban GA (1985) The effect of practice on the oblique effect in line orientation judgments. Vision Research 25: 1679–1687.
13. Karni A, Sagi D (1991) Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. Proceedings of the National Academy of Sciences 88: 4966–4970.

14. Li W, Piech V, Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. Nature Neuroscience 7: 651–657.
15. Ahissar M, Hochstein S (1996) Learning pop-out detection: specificities to stimulus characteristics. Vision Research 36: 3487–3500.
16. Schoups A, Orban GA (1996) Interocular transfer in perceptual learning of a pop-out discrimination task. Proceedings of the National Academy of Sciences 93: 7358–7362.
17. Schoups A, Vogels R, Qian N, Orban G (2001) Practising orientation identification improves orientation coding in v1 neurons. Nature 412: 549–553.
18. Ghose G, Yang T, Maunsell J (2002) Physiological correlates of perceptual learning in monkey V1 and V2. Journal of Neurophysiology 87: 1867–1888.
19. Dosher BA, Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proceedings of the National Academy of Sciences 95: 13988–13993.
20. Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends In Cognitive Science 8: 457–464.
21. Law C, Gold JI (2008) Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nature Neuroscience 11: 505–513.
22. Yang T, Maunsell J (2004) The effect of perceptual learning on neuronal responses in monkey visual area V4. Journal of Neuroscience 24: 1617–1626.
23. Raiguel S, Vogels R, Mysore SG, Orban GA (2006) Learning to see the difference specifically alters the most informative v4 neurons. Journal of Neuroscience 26: 6589–6602.
24. Fahle M (2005) Perceptual learning: specificity versus generalization. Current Opinion in Neurobiology 15: 154–160.
25. Sireteanu R, Rettenbach R (1995) Perceptual learning in visual search: fast, enduring, but non-specific. Vision Research 35: 2037–2043.
26. Logan G (1988) Toward an instance theory of automatization. Psychological review 95: 492–527.
27. Shiffrin R, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological review 84: 127–190.
28. Goldstone RL (1998) Perceptual learning. Annual Review of Psychology 49: 585–612.
29. Navalpakkam V, Itti L (2006) Top-down attention selection is fine grained. Journal of Vision 6: 1180–1193.
30. Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition 12: 1093–1123.
31. Motter BC, Belky EJ (1998) The guidance of eye movements during active visual search. Vision Research 38: 1805–1815.
32. Bichot NP, Schall JD (1999) Effects of similarity and history on neural mechanisms of visual selection. Nature Neuroscience 2: 549–554.
33. Rutishauser U, Koch C (2007) Probabilistic modeling of eye movement data during conjunction search via feature-based attention. Journal of Vision 7: 5.
34. Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognitive Psychology 12: 97–136.
35. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 20: 6.
36. Navalpakkam V, Itti L (2007) Search goal tunes visual features optimally. Neuron 53: 605–617.
37. Wolfe J (1994) Guided search 2. 0. A revised model of visual search. Psychonomic Bulletin & Review 1: 202–238.
38. Hillyard SA, Vogel EK, Luck SJ (1998) Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philosophical Transactions of the Royal Society B: Biological Sciences 353: 1257–1270.
39. Phillips MH, Edelman JA (2008) The dependence of visual scanning performance on saccade, fixation, and perceptual metrics. Vision Research 48: 926–936.
40. Hooge IT, Erkelens CJ (1999) Peripheral vision and oculomotor control during visual search. Vision Research 39: 1567–1575.
41. Luria SM, Strauss MS (1975) Eye movements during search for coded and uncoded targets. Perception & Psychophysics 17: 303–308.
42. Zelinsky GJ (1996) Using eye saccades to assess the selectivity of search movements. Vision Research 36: 2177–2187.
43. Engel FL (1971) Visual conspicuity, directed attention and retinal locus. Vision Research 11: 563–576.
44. Courtney AJ, Chan HS (1986) Visual lobe dimensions and search performance for targets on a competing homogeneous background. Perception & Psychophysics 40: 39–44.
45. Findlay J, Gilchrist I (2003) Active vision: The psychology of looking and seeing Oxford University Press Oxford.
46. Motter BC (1994) Neural correlates of feature selective memory and pop-out in extrastriate area v4. Journal of Neuroscience 14: 2190–2199.
47. Motter BC (1994) Neural correlates of attentive selection for color or luminance in extrastriate area v4. Journal of Neuroscience 14: 2178–2189.
48. Wolfe JM, Horowitz TS (2004) What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5: 495–501.
49. Herzog M, Fahle M (1998) Modeling perceptual learning: Difficulties and how they can be overcome. Biological Cybernetics 78: 107–117.

Review

Cell PRESS

*Feature Review*

# Mechanisms of top-down attention

**Farhan Baluch[1] and Laurent Itti[1,2]**

[1] Neuroscience Graduate Program, University of Southern California, Los Angeles, CA, USA
[2] Department of Computer Science, University of Southern California, Los Angeles, CA, USA

**Attention exhibits characteristic neural signatures in brain regions that process sensory signals. An important area of future research is to understand the nature of top-down signals that facilitate attentional guidance towards behaviorally relevant locations and features. In this review, we discuss recent studies that have made progress towards understanding: (i) the brain structures and circuits involved in attentional allocation; (ii) top-down attention pathways, particularly as elucidated by microstimulation and lesion studies; (iii) top-down modulatory influences involving subcortical structures and reward systems; (iv) plausible substrates and embodiments of top-down signals; and (v) information processing and theoretical constraints that might be helpful in guiding future experiments. Understanding top-down attention is crucial for elucidating the mechanisms by which we can filter sensory information to pay attention to the most behaviorally relevant events.**

## Introduction

Language is infused with idiomatic expressions that make explicit the distinction between bottom-up (BU) and top-down (TD) processes of attention. We might ask someone to 'pay attention to the road' while driving, which implies a voluntary choice to allocate resources to a subset of the perceptual input. Alternatively, we might remark that the orange sports car really 'caught our attention'. In this case, the resource has been involuntarily captured rather than voluntarily allocated. The distinction is not limited to idiomatic expressions, but rather stems from disparate modes of attentional processing [1]. BU attention is deployed very rapidly and depends exclusively on the properties of a sensory stimulus. By contrast, TD attention is slower and requires more effort to engage.

In the modality of vision, the two modes (BU and TD) give rise to the psychophysical phenomenon of pop-out and set-size effects. In a typical visual search experiment, a subject is presented with a number of items on a display and is asked to find a target item within this display, such as a bar with a particular orientation, or color, or a combination of the two. Pop-out occurs when the target item is significantly distinct from the surrounding items (distractors), such as a horizontal bar among several vertical bars. This different item automatically attracts BU attention (or pops-out) rapidly and independently of the number of distractors [2,3]. By contrast, when the target item is distinguished only by taking into account the conjunction of its features, such as color and orientation, BU cues alone cannot efficiently guide attention and TD attention must be recruited to scan the display. This gives rise to search times that increase with the number of distractors; in other words, a set-size effect is observed. In most real-life situations, the responses of the nervous system to a sensory input depend on both BU influences driven by the sensory stimulus and TD influences shaped by extra-retinal factors such as the current state and goal of the organism [4,5].

A distinction is also made between two types of TD mechanisms. The first type is intuitively associated with TD and is called the volitional TD process, which can exert its influence through acts of will. The second type is known as a mandatory TD process and it is an automatic, percept-modifying TD mechanism that is pervasive and that volition cannot completely eliminate. The latter TD process can develop through experience-dependent plasticity or during development, and includes contextual modulation

## Glossary

**BU influence**: influence on the nervous system due to extrinsic properties of the stimuli.
**Conjunction search**: search task in which a subject is required to find a target item among several distractors, and the target is defined by a unique conjunction of features. In this type of search task, locating the target is more difficult because distractors share some of the features of the target and thus the target does not obviously stand or pop out.
**Covert attention**: attention paid to a subset of the sensory inputs through mental focusing.
**Feed-forward sweep**: first epoch of neural activity that travels from lower to higher visual areas on the onset of a visual stimulus via feed-forward connections.
**Mandatory TD process**: attentional process that influences sensory processing in an automatic and persistent manner.
**Overt attention**: attention paid through orienting of sensory organs toward a sensory input of interest.
**Percept**: mental impression of an external stimulus.
**Pop-out search**: search task in which a subject is required to find a target item among several distractors, and the target is defined by a unique visual feature not shared with any of the distractors. The target thus stands or pops out and is easy to find.
**Priority map**: map of visual space constructed from a combination of properties of the external stimuli, and intrinsic expectations, knowledge and current behavioral goals.
**Recurrent epoch**: second epoch of neural activity that occurs after an initial response to onset of a stimulus and is mediated by intra-cortical horizontal connections and inter-cortical feedback connections.
**Saliency map**: map of stimulus conspicuity over visual space.
**Set-size effect**: in search tasks, a set-size effect is observed if the time required to find the target depends on the total number of items in the display (the set size).
**Task-relevance map**: map of behaviorally relevant locations over visual space.
**TD influence**: influence on the nervous system due to extra-retinal effects such as intrinsic expectations, knowledge and goals.
**Volitional TD process**: attentional process that exerts influence on sensory processing through an act of volition, such as willfully shifting attention to the right part of space.

*Corresponding author:* Itti, L. (itti@usc.edu).

 0166-2236/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.tins.2011.02.003 Trends in Neurosciences, April 2011, Vol. 34, No. 4

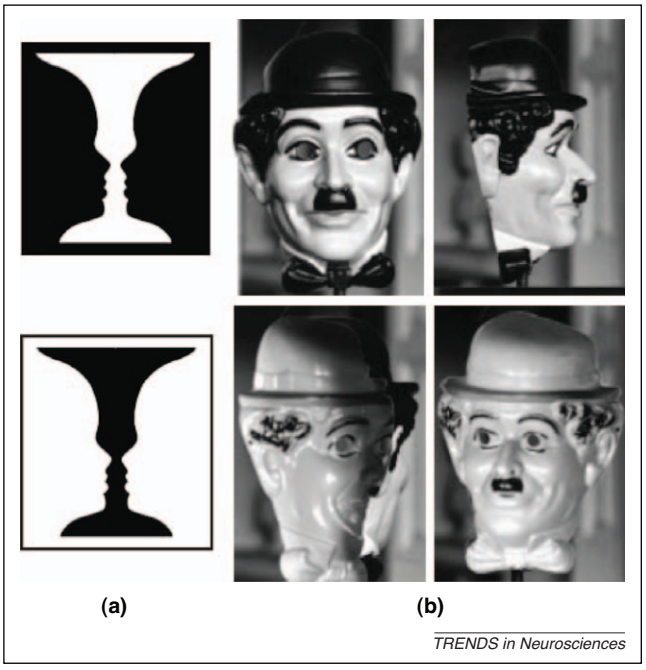



**Figure 1**. Mandatory versus volitional TD processes. **(a)** Rubin's vase illusion is an example of a volitional TD process. The percept can be switched from face to vase and vice versa through act of will, demonstrating TD modulation of perceptual processing in a volitional and dynamic manner. **(b)** Four frames from a demonstration of a rotating mask that seems to be convex, even in places where it is in fact concave. At the start of the rotation (top two frames), the mask is convex and, as it rotates, the viewer begins to see the inside of the mask (bottom two frames), but this still seems to be convex. This demonstrates an inherent bias in perceiving faces as convex rather than concave, even when this contradicts BU sensory information, and thus provides an example of mandatory TD processing. Reproduced with permission from [134].

of perception [5,6]. A striking example of the dichotomy between these two mechanisms is presented in Figure 1.

Previous work has extensively studied the effects of TD attention on target brain regions, including modulatory effects in early sensory areas [5,7,8]. Significant progress has been made in isolating the possible sources of TD signals [9], especially within the now well-studied fronto-parietal attention network [10]. Much less understood at present are the exact pathways, contents, meaning and form of the signals that are sent from the top down. Here, we review recent findings from physiology, lesion and computational studies that have attempted to elucidate the mechanisms and signals involved in TD modulation of sensory processing. To focus this review, we mainly concern ourselves with visual perception and the volitional TD process, although similar principles can apply in other modalities.

## Brain structures and circuits of visual attention

Visual processing begins in the retina, which sends parallel streams of information to the brain through its diverse set of retinal ganglion cells and their unique interactions within the retinal circuitry [11]. A majority of retinal projections reach the lateral geniculate nucleus (LGN) and a much smaller number (approx. 10%) connect to the superior colliculus (SC). The LGN sends projections to the primary visual cortex (V1), the initial site of processing in the cortical feed-forward visual pathway. This pathway has been functionally divided into the dorsal and ventral streams [12]. The dorsal stream has been described as the 'where' pathway and leads from area V1 to motion processing areas [medial temporal (MT) and medial superior temporal (MST)] and parietal cortices. The ventral or 'what' pathway comprises striate (V1) and extrastriate areas (V2, V3, V4) and leads to the inferotemporal cortex (IT), believed to be the last feature-selective area in the visual processing hierarchy.

Modulatory effects of attention have been observed in the constituent structures of both the dorsal and ventral streams. The first structure subject to strong attention effects is the SC. The SC is a layered midbrain structure that receives direct input from the retina, as well as feedback inputs from area V1. Salient visual events are represented in the superficial layers of the SC [13,14] and can further combine, in the deeper layers, with TD information to give rise to a priority map that guides attention [14]. This attention map is probably shared or jointly computed with the lateral intraparietal (LIP) region of the cortex [15], the frontal eye fields (FEF) [16] and visual cortices, through direct afferent connections from the cortex to the SC, as well as indirect efferent connections from the SC to the cortex via the pulvinar [17]. These connections are important for communicating attention-related signals to higher cortical areas while bypassing the canonical ventral pathway.

Situated a level above the SC in the visual processing hierarchy, are the thalamic nuclei, which are involved in processing many types of sensory information and are susceptible to modulation by attention. The LGN is the most visually responsive of the thalamic nuclei, and both physiological studies in monkeys and imaging studies in humans have shown that attention can modulate signals in the LGN [18,19]. The modulation includes enhancement of neural responses to attended stimuli and suppression of unattended stimuli [19]. Thus, visual sensory information is already subject to attentional modulation even before entering the cortex.

The first cortical stage of visual processing, area V1, is the first major feature-sensitive area of processing and is also modulated by attention. However, these effects are relatively weak [20,21]. Moving up the visual processing hierarchy from V1, V2, V4 to IT, receptive field sizes increase and visual areas are progressively more sensitive to features than spatial locations of stimuli. When attention is allocated to a certain part of visual space, neurons encoding this part are facilitated (a phenomenon known as spatial attention). The allocation of attention to a particular non-spatial feature, such as the color or orientation of an object, facilitates neurons encoding the attended feature (feature-based attention). Along the ventral pathway, extrastriate areas V4 and IT have large receptive fields and effects of feature-based attentional modulation are more evident. Motion-sensitive MT and MST areas are also modulated by both spatial and feature-based attention [22]. This tendency for combined modulation of sensory signals by both spatial and feature-based attention increases as the signals progress from lower to higher cortical areas such as the LIP.

The LIP area has been studied extensively and several excellent recent reviews have described its diverse roles in

attention, reward, and occulomotor behavior [15,23,24]. It is important to point out that responses in area LIP can be driven by both BU factors, such as stimulus salience, and TD factors, such has behavioral relevance of stimuli [25], the current locus of attention [26] and occulomotor planning [15]. Therefore, the LIP is another candidate structure (beyond the SC described above) where BU and TD influences can combine to give rise to a spatial priority map [15]. The many facets of observed responses in the LIP can be attributed to the fact that both BU and a diverse set of TD influences can give rise to behavioral priority, and thus modulate LIP responses, which suggests that the LIP encodes priority in a manner largely agnostic to the factors that caused the priority [15]. Through direct feedback connections [27] or connections via the pulvinar to visual areas (see below), the LIP can communicate the fused signals to other brain areas for biasing or further attentional processing.

FEF neurons also represent salient stimuli, specifically stimuli that vary significantly from surrounding items in a visual display (known as odd-ball stimuli). The FEF has also been described as a region with neural responses characteristic of a priority map [16]. Single-unit responses in monkey FEF exhibit transients on stimulus onset, followed by a later response (latency of ~100 ms) that discriminates an odd-ball stimulus from surrounding distractors [14,16]. This suggests that the FEF computes salience in the recurrent epoch rather than the initial feed-forward sweep [28,29]. The FEF's connections to motor neurons in intermediate and deep layers of the SC make it an important structure in occulomotor behaviors associated with attention. In addition to this role of the FEF in representing BU salience, we examine in the following section its involvement in projecting TD signals to other regions of the attentional network.

Effects of attention have also been observed in prefrontal cortex (PFC). The PFC is thought to be involved in short-term memory processes, and recent studies suggest that the PFC also exhibits strong attentional selection related signals [30,31]. Owing to its involvement in short-term memory and its position high in the visual hierarchy, it is also the primary candidate for generating TD signals and sending them to sensory cortex for spatial or feature-based attentional biasing.

Therefore, the LGN, the striate and extrastriate cortex (areas V1, V4, IT and MT), as well as the SC, pulvinar, LIP, FEF and PFC, are known to be involved in attentional processes. Modulatory attentional signals are found as early as in the SC (a brainstem structure) and in the LGN, the first stop along the visual processing hierarchy [18,19]. These signals act progressively sooner and with stronger modulatory power going up from area V1 to area

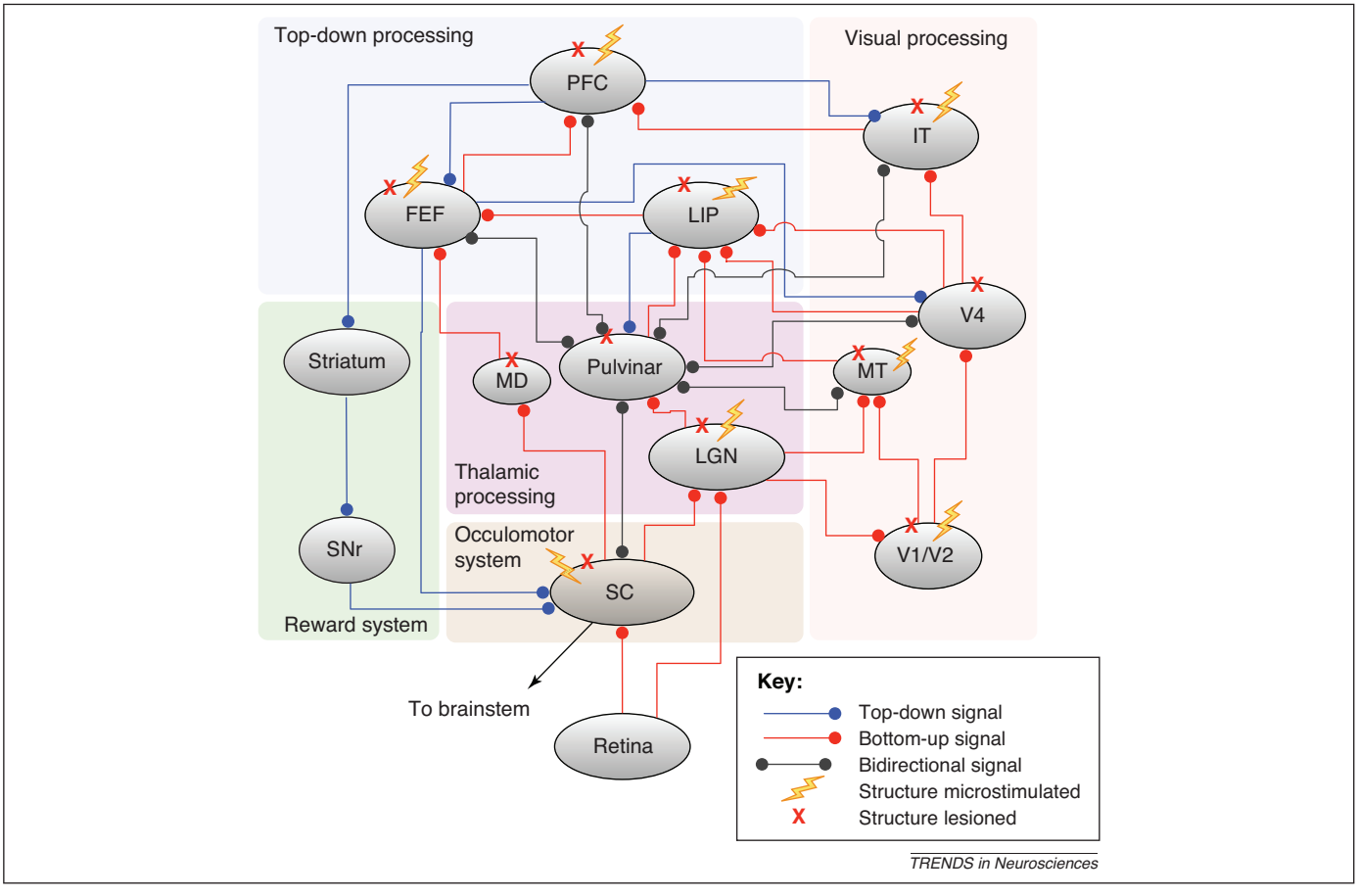



**Figure 2**. Flow of attentional signals in brain structures that have been implicated as being involved in attentional studies. The flash symbol indicates that a candidate structure has been microstimulated and an X indicates that the structure has been lesioned in a previous study (see Table 1 for details). The connections show the most likely type of signal being transmitted between two areas; TD signals are shown in blue, BU signals in red and bidirectional signals in gray. Abbreviations: SC, superior colliculus; SNr, substantia nigra pars reticulata; MD, mediodorsal thalamus; LGN, lateral geniculate nucleus; IT, inferotemporal cortex; MT, middle temporal area; LIP, lateral intraparietal area; FEF, frontal eye fields; PFC, prefrontal cortex.

IT [20]. These signals can bias attention for particular visual locations [32], visual features [33–36], or both. The characteristic signature of these attentional modulations onto target sensory areas includes heightened gain, sharpened tuning and other end-effects, as reviewed previously [8,37,38]. In the following section, we examine the areas that are specifically involved in mediating TD attentional signals.

## Pathways of TD attention

In this section, we focus on lesion and electrophysiological studies, particularly those using methods of microstimulation and simultaneous recordings in the brain areas identified in the previous section. These areas form an attentional network (Figure 2) and we consider how TD information is relayed in this network. Microstimulation, together with reversible inactivation [using either pharmacological agents such as muscimol or transcranial magnetic stimulation (TMS)] and permanent lesion studies, have enabled researchers to go from correlation to causation in the study of perception and attention (Table 1).

It has been suggested that all sensory stimuli compete for entry into working memory [39]. Working memory not only stores information, but also enhances this information and actively generates TD attentional signals that bias feature-sensitive brain regions, and is thus vital for accomplishing behavioral goals [39]. An elegant study demonstrated that the PFC transmits the contents of working memory to the visual system by using a posterior-split-brain paradigm [40]. In this study, monkeys were presented with a visual cue in either the left or right hemifield, followed by a probe stimulus. The task was to respond to the appearance of the probe that had previously been associated with the cued item. BU signals were recorded by presenting the cue in the hemifield ipsilateral to the recording site in the IT (i.e. direct BU path from the retina up to IT), whereas TD signals could be recorded from area IT by presenting the cue in the visual hemifield contralateral to the recording site in the IT. The posterior callosum transection precluded direct communication between visual cortices from both sides of the brain, so it was hypothesized that the TD signals were fed back from the PFC to area IT (Figure 3a). To move to a more causal explanation, the next experiment involved transection of the anterior corpus callosum (thereby cutting that hypothetical pathway), which resulted in a lack of response from the IT cells [40]. These results demonstrated that TD signals correlating with working memory emanate from the PFC and feed back into the ventral stream. A more recent study also used the posterior-split-brain paradigm in conjunction with unilateral PFC removal and demonstrated that performance on a search task was mainly impaired when the goal of the search was switched on a regular basis [41]. This study thus highlighted the importance of the PFC in switching the TD context. It has also been found that microstimulation of the PFC leads to biases in target selection towards or away from the stimulation field, which demonstrates how TD signals can affect occulomotor behavior [42]. Furthermore, the sheer connectedness of the PFC suggests that its effects are pervasive and are driven by a combination of goals, rewards, salience, and planning of motor actions [9,39].

The next area proximal to the PFC, and an important player in TD attention, is the FEF. Sub-threshold FEF stimulation enhances responses of V4 neurons in the presence of a stimulus in their receptive field (Figure 4a) [43]. This demonstrates that descending TD signals from the FEF bias processing in area V4. These results were replicated in analogous regions of the barn owl [44]. The comparison of local field potentials (LFP, which may be strongly driven by afferent inputs from other brain regions) and spiking activity in the FEF (which represents intrinsic activity of FEF neurons) revealed that target-selective signals appeared in spiking activity before showing a difference in the LFP, which suggests that spatial selection was computed locally in the FEF [29]. There is speculation that this emergence of selection is communicated down to ventral regions through a synchronization of gamma-band activity between the FEF and area V4 [45]. However, a lesion study demonstrated that temporary inactivation of the FEF (using a GABA-A receptor agonist, muscimol) led to deficits not only in visually guided saccades, but also in shifts of attention during either pop-out or conjunction visual searches [46]. Contrary to an earlier study [29], these findings suggested that the FEF, although involved in covert attention, does not locally compute the selection but is rather a participant in a network with heavy involvement of the LIP.

Area LIP is strongly connected to the FEF and is integral to the attentional network through both anatomical and functional characterization. Suprathreshold microstimulation in the posterior parietal cortex (PPC), which includes both area LIP and the ventral intraparietal area (VIP), induces saccades; however, the current required to induce saccades is significantly higher compared to that required when microstimulating the FEF, which suggests that the connection from the PPC to the occulomotor system might not be a direct one. Subthreshold stimulation results in a shift of covert attention [47]. Interestingly, a non-spatial effect was also found whereby reaction times in detecting a target decreased irrespective of whether a valid, invalid, or no cue was presented [47]. This suggested that microstimulation of the LIP can override the cue signal and orient attention to the visual location corresponding to the site of stimulation. Evidence from lesion studies demonstrates that damage or inactivation of the LIP causes deficits only in the presence of multiple stimuli [48,49]. These results point to an additional role of the LIP in resolving competition among stimuli represented at lower levels through TD connections to these levels [7,50].

The aforementioned studies did not, however, differentiate between the dorsal (LIPd) and ventral (LIPv) subdivisions of the LIP. In a more recent study, the effects of local reversible inactivation (using a GABA-A receptor agonist) in areas LIPd and LIPv have been studied separately [51]. Interestingly, the many dimensions of LIP responses demonstrated previously [23] were shown to reside in disparate subdivisions of the LIP. Inactivation of the LIPd affected performance on simple saccade tasks but left visual search intact, whereas temporary lesions of the LIPv led to deficits in both search and saccadic performance [51]. The authors stressed that deficits in saccadic performance after LIPd inactivation were far smaller than

**Table 1. Microstimulation and lesion studies of different brain structures involved in attention.[a,b]**

| Brain region | Microstimulation studies | Refs | Lesion studies | Refs |
|---|---|---|---|---|
| | ***Implications for attentional processing*** | | ***Implications for attentional processing*** | |
| SC | Shift of spatial attention | [55] | Deficit in target selection | [58,68] |
| | Perceptual facilitation at site of stimulation | [56] | Deficit in perceptual decision in presence of distractors | [60] |
| | Selection of target independent of motor plan | [57] | | |
| | Signal transmitted to MT via Pulvinar | [70] | | |
| LGN | Elicits visual percepts | [111] | Eliminates residual visual responses in extrastriate cortex after V1 lesion | [112] |
| | | | Disruption of smooth pursuit eye movements | [113] |
| | | | Deficits in target detection (human) | [110] |
| Pulvinar | – | | No deficit in saccadic behavior | [67] |
| | | | No deficit in visual search | [68] |
| | | | Deficit in suppression of distractors during search (human) | [65] |
| | | | Spatial and temporal attention deficits with anterior and posterior lesions respectively (human) | [66] |
| V1 | Target selection disrupted with upper layer stimulation, facilitated with lower layer stimulation | [114] | Deficit in motion detection and discrimination | [117] |
| | Lower current thresholds needed for evoking saccades in lower layers | [115] | Deficit in saccade targeting | [118] |
| | Median current of 5.2 μA (6.6 μA) required for behavioral detection of stimulation[c] | [116] | | |
| V4 | – | | Deficit in distractor suppression when target and distractor are inside RF of neuron | [119] |
| | | | Deficit in distractor suppression | [133] |
| IT/TE | Biases perceptual judgement in visual classification | [53] | No behavioral deficit when lesion is made in infantile monkeys | [120] |
| | Bias in selection of stimulus category | [54] | Deficit in distractor suppression | [119] |
| | Median current of 10.3 μA (11.3 μA) required for behavioral detection of stimulation[c] | [116] | | |
| MT | Bias in motion direction discrimination | [122] | Loss in perception of motion | [124] |
| | Bias in motion direction during stimulus presentation but not during memorizing period | [123] | Loss in perception of motion more evident in noisy conditions | [132] |
| | Median current of 10.1 μA required for behavioral detection of stimulation[c] | [116] | | |
| LIP | Sub and suprathreshold stimulation lead to covert and over shifts of attention respectively | [47] | Deficit in distractor suppression even when stimuli are non overlapping within RF, contrast with [121] | [49] |
| | Bias in visual selection | [125] | Dorsal lesion leads to occulomotor deficits ventral lesion leads to attentional and occulomotor deficit | [51] |
| | | | Affects performance in tasks requiring spatial attention | [48] |
| FEF | Enhanced response elicited in V4 | [43] | Deficit in target detection | [46] |
| | Facilitation akin to allocation of covert attention | [126] | Enhanced contrast sensitivity in fovea but not periphery (human) | [128] |
| | Bias toward direction of saccade plan rather than location of attention | [127] | Disruption of facilitation by saccade plan to location corresponding with stimulation site (human) | [129] |
| PFC | Bias in target selection | [42] | Loss of TD signal recorded in IT | [40] |
| | Disruption in saccadic activity | [130] | Decrease in behavioral performance when cue is frequently switched | [41] |
| | | | Elimination of acetylcholine release in sensory cortex after stimulus presentation (rat) | [131] |

[a]All studies have been conducted in monkeys unless otherwise denoted.

[b]RF, receptive field; SC, superior colliculus; LGN, lateral geniculate nucleus; IT, inferotemporal cortex; MT, middle temporal area; LIP, lateral intraparietal area; FEF, frontal eye fields; PFC, prefrontal cortex.

[c]Stimulation current values reported in two monkeys (see [116] for details).

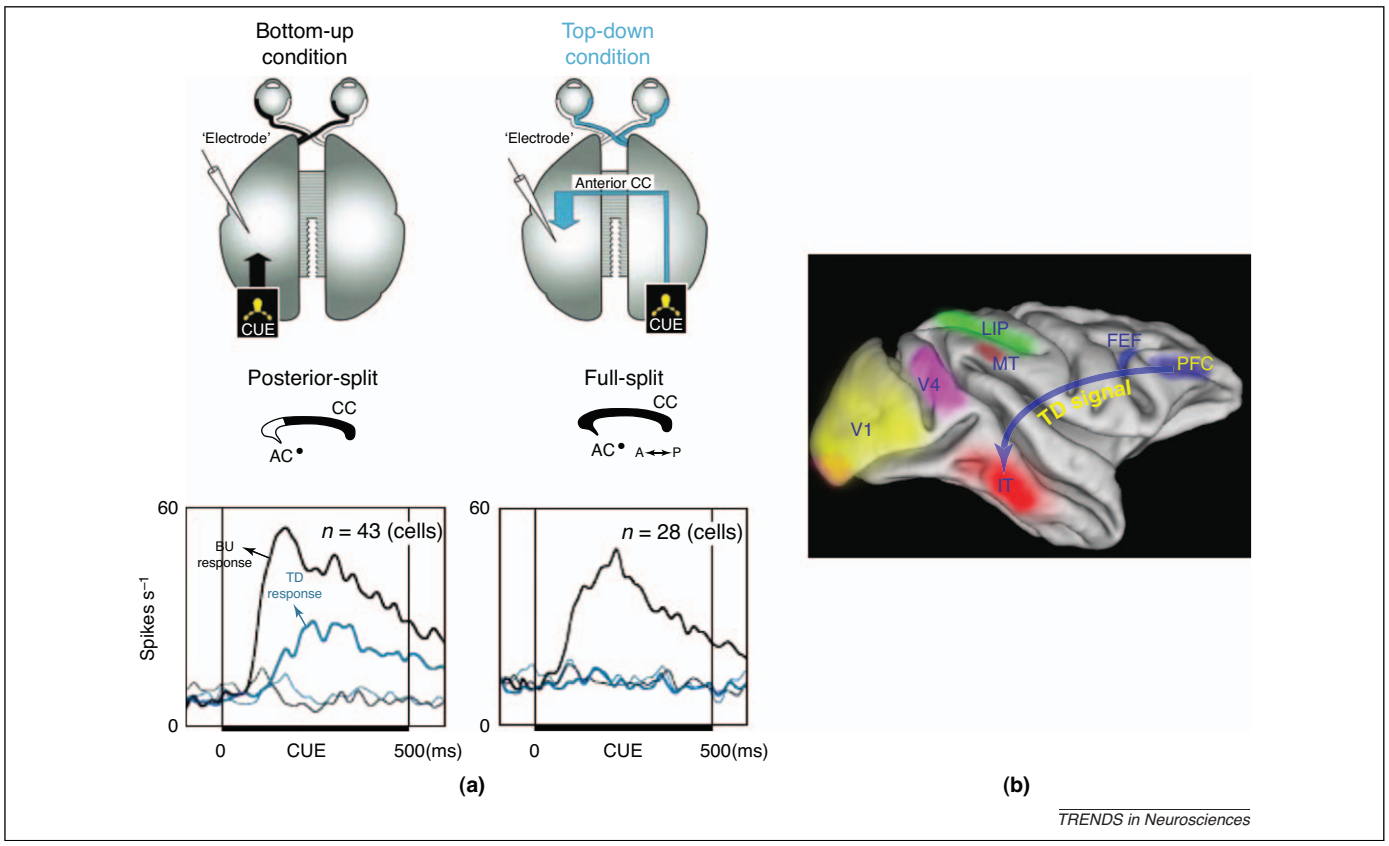



**Figure 3**. Role of the PFC in mediating TD attentional signals. **(a)** Posterior-split-brain paradigm. In this study, monkeys had to associate stimulus A with stimulus B [40]. Stimulus A was then presented as the cue, followed by a probe stimulus, and the task was to release a lever when the probe matched the associated stimulus B. Neurons in the inferotemporal cortex (IT) were recorded in one hemisphere while the cue was presented either contralateral to the recording site (BU condition, i.e. information about the cue could reach area IT directly; top-left panel) or ipsilateral (TD condition, i.e. information about the cue could only reach area IT via the anterior corpus callosum; top-right panel). The bottom panels show neural responses in BU (black trace) and TD (blue trace) conditions. The left-hand plot shows the responses after a posterior split, demonstrating how cue information could reach area IT in both the BU and TD conditions. The right-hand plot shows complete abolition of the TD signal after a full split of the corpus callosum (CC). This is one of the clearest demonstrations of the two different types of signal, BU and TD, recorded in a visual area. Reproduced with permission from [40]. **(b)** 3D rendering of macaque monkey brain showing regions involved in visual processing and TD attention. The areas include the first visual area (V1), fourth visual area (V4), medial temporal cortex (MT), lateral intraparietal cortex (LIP), frontal eye fields (FEF), inferotemporal cortex (IT) and prefrontal cortex (PFC). The blue arrow shows the pathway for TD signals investigated in the experiment shown in (a). Rendering of the brain was done using a macaque atlas data set [135] processed using the Caret software [136].

those observed after inactivation of the FEF and SC. They thus concluded that the LIP might influence or modulate the motor decision but that the final decision is made by more downstream structures such as the SC and FEF. This coincides with the view that attentional selection might indeed be separate from motor selection [23]. As discussed previously, the diversity of properties exhibited by LIP neurons might reflect the fact that it encodes priority without regard for what caused the priority, BU or TD influences.

We now consider feature-selective visual areas V1, V2, V4, MT and IT. These visual processing areas drive BU attentional signals and are targets for TD attentional biasing signals. For accurate biasing of sensory signals, specific local circuitry and the nature and size of receptive fields in each of these areas must constrain the nature and granularity of TD signals. Two types of feedback signals from higher cortical regions or thalamus can influence the visual processing areas [52]. One type of feedback signal can flow between a higher visual processing area to a lower one within the visual processing hierarchy (Figure 4b). Another type of feedback signal can flow between an attention area such as the FEF and a processing area such as area V4. Figure 4a presents data from a study that demonstrates a specific example of this type of feedback signal [43]. The flow of TD attentional signals from the PFC to area IT is another example of how TD attentional signals from higher cortex can influence a feature sensitive sensory area [40]. Microstimulation in area IT results in biases of object recognition [53], or even of face detection when microstimulating face-selective sites within area IT [54]. The striate and extrastriate cortices, therefore, are all amenable to modulation by TD attention through feedback connections from higher to lower visual areas.

In summary, TD signals can emerge from the PFC to bias visual cortices through direct connections, such as from the PFC to area IT, or possibly through the pulvinar (see below). Similarly, there is evidence that a direct connection from the FEF to area V4 might exist, which further demonstrates the possible communication of TD information from higher cortex to sensory areas. TD signals from the PFC probably contain detailed information about the target and this information might be used to bias feature-selective areas of sensory cortex. The FEF and LIP,

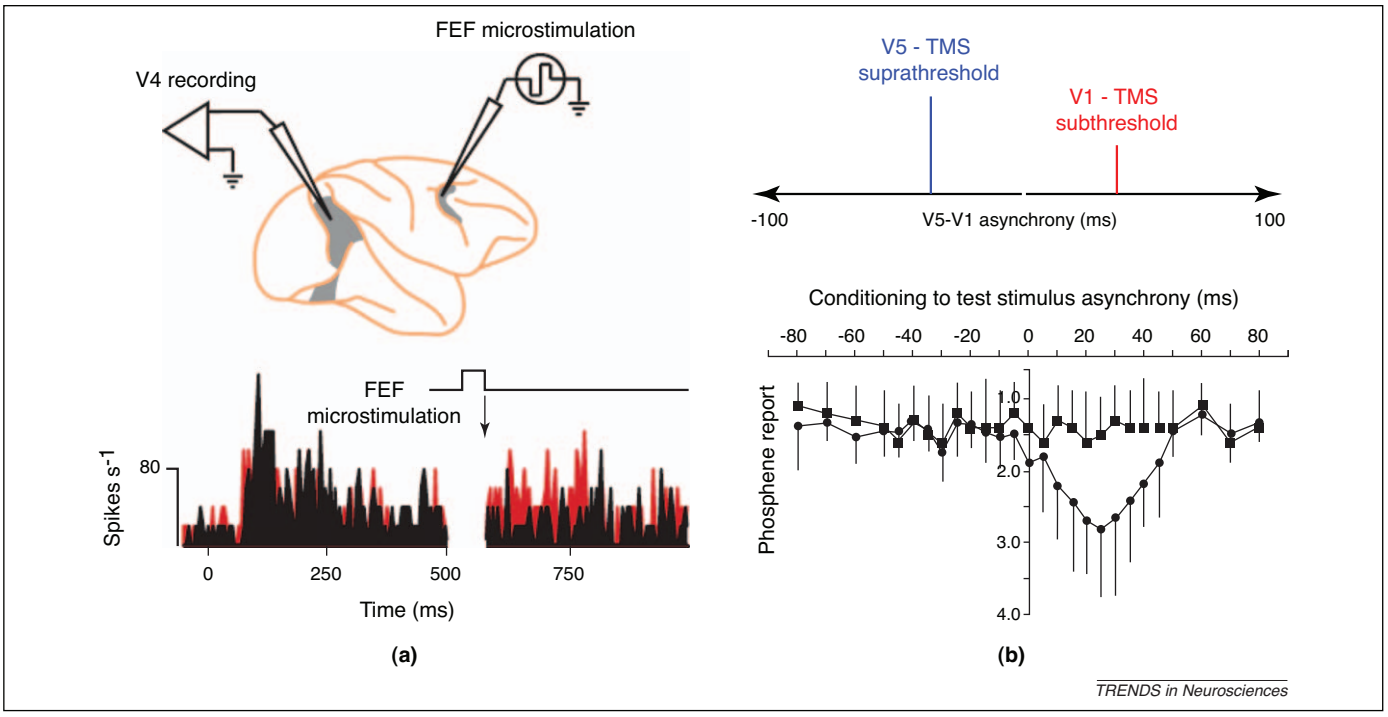



**Figure 4**. Role of feedback from higher to lower cortical areas in mediating attention and perception. **(a)** Neuronal activity from visual area V4 was recorded in monkeys simultaneously as the frontal eye field (FEF) was microstimulated (top panel). Histogram of neuronal activity in area V4 (bottom panel) in the control condition (black) and the stimulation condition (red). Clear enhancement of the response is evident after FEF stimulation. This demonstrates the role of frontal areas in modulating responses in a sensory visual area such as area V4. Reproduced with permission from [43]. **(b)** Visual area V5 in human subjects was stimulated with suprathreshold transcranial magnetic stimulation (TMS) pulses, followed by subthreshold TMS stimulation of visual area V1 [137]. The top panel shows the TMS paradigm used. The bottom panel shows a plot of subjective report by human subjects of phosphene perception resulting from TMS stimulation as a function of time lag between V1 and V5 stimulations (negative X values correspond to area V1 stimulation before area V5). A Y-value of 1.0 indicates that the subject perceived that a phosphene was present and moving; a value of 2.0 indicates that a phosphene was present but the subject was uncertain of motion; and a value of 3.0 indicates that the subject could see the phosphene but it was stationary. Results show that disruption of V1 activity between 5 and 45 ms after V5 stimulation results in the absence of motion, which thus demonstrates the importance of feedback signals to early visual areas for the perception of motion. Reproduced with permission from [137].

in particular, might host spatial maps encoding the behavioral relevance of visual space dependent on both BU and TD factors.

## Subcortical influences on TD attention

Evidence suggesting that cortical areas have a strong influence on attention was discussed in the previous section. There are also several subcortical areas that play a crucial role in defining and communicating attentional signals (Figure 5a). It has been demonstrated that the phenomenon of change blindness, in which changes to a particular part of a visual scene go undetected, could be eliminated in monkeys by placing an attention-grabbing salient stimulus in the location where the blindness occurs [55]. Interestingly, the same effects were also observed by microstimulating the SC where receptive fields overlapped with the region of blindness [55]. This demonstrated that stimulation of the SC is equivalent to adding salience to a region of space; in other words, the SC can strongly bias attentional deployment. Another study demonstrated enhanced behavioral performance on a perceptual task with stimuli at locations corresponding to the site of stimulation in the SC [56], mimicking the effects of a shift of attention.

In another study, microstimulation of the SC in monkeys led to a bias in target selection decisions [57], which demonstrates that the SC is also involved in target selection. Conversely, inactivation of the SC led to target selection errors [58]. The SC is therefore involved in both attentional selection and saccadic behavior. One study was able to elegantly dissociate saccade preparation signals from attentional signals [59], which clarified any ambiguity about the dual roles of the SC in occulomotor behavior and attentional control. This study involved recording from visual, visuomotor, and motor neurons in monkey SC. This revealed that visuomotor neurons encode the shift of covert attention (Figure 5b). It has also been shown that the SC is involved in gating covert attention signals used for making perceptual decisions by higher cortical areas [60].

The SC connections to the FEF and LIP, together with its role as an occulomotor structure, make it an important structure in mediating covert and overt attention. Furthermore, given its direct involvement in occulomotor behavior, it has been suggested that the SC could host the final priority map that guides attention based on a fusion of TD and BU attentional signals received from cortex and elsewhere [14].

Moving up the neuraxis to the thalamus, three important nuclei associated with visual functions are found: the LGN, the thalamic reticular nucleus (TRN) and the pulvinar nucleus. The LGN and TRN modulate their signals in a reciprocal manner (Figure 5c). When monkeys attended inside the receptive field of the recorded TRN neuron, the responses of this cell were reduced, whereas responses in the LGN were enhanced [18]. This reciprocal response in the TRN and LGN neurons was found in the initial phase of

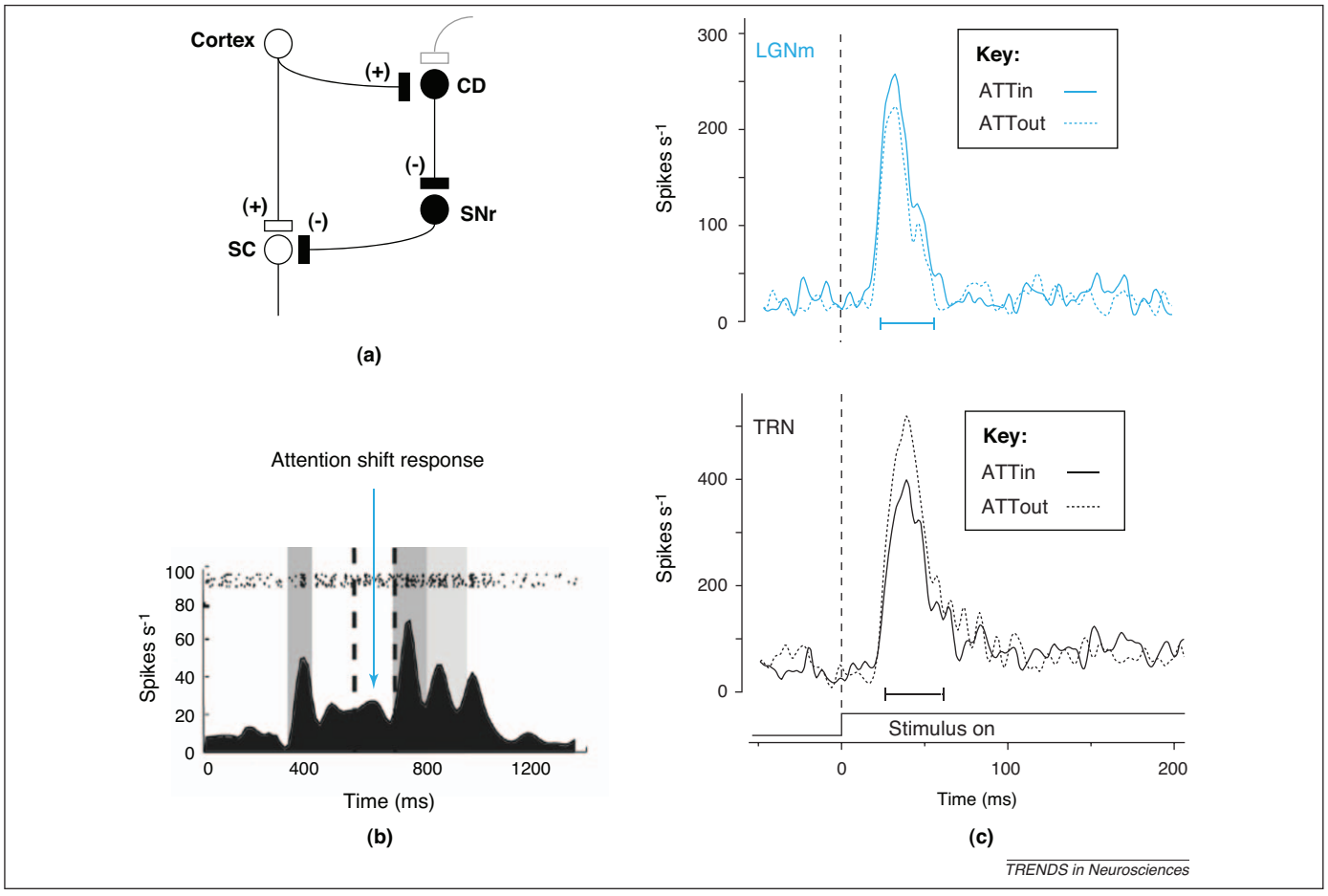



**Figure 5.** Role of subcortical structures in attention. **(a)** Schematic drawing of circuitry that has been proposed to be involved in the generation of eye movements towards locations of reward ([78]). The cortex sends excitatory inputs to both the superior colliculus (SC) and the caudate nucleus (CD). The CD in turn inhibits the substantia nigra pars reticulata (SNr), which then reduces its tonic inhibition on the SC. A disinhibited SC enables eye movements to be made. Reproduced with permission from [78]. **(b)** Neuronal activity from a visuomotor cell in the SC. Monkeys were first presented with a spatial cue, followed by an oriented stimulus at the cued location. Monkeys then made a saccade in the direction corresponding to the orientation of the stimulus. The orientation was always orthogonal to the location of the cue, and this dissociates shifts of attention from saccadic behavior. The plot shows responses of a visuomotor SC cell, which shows significant activity in the attention shift period (between the dashed lines) that occurs immediately after presentation of the cue, whereas purely motor cells in the deeper layers of the SC did not show such a response (data not shown). Reproduced with permission from [59]. **(c)** Neuronal activity recorded from the thalamus in awake behaving monkeys. The monkeys were presented with a central cue that instructed them to attend to one of two peripheral oriented bar stimuli, one inside the receptive field (RF) of a recorded neuron and one outside the RF. The top shows the spike density of a magnocellular lateral geniculate nucleus (LGNm) neuron that exhibits an enhanced response when the monkey attends to a stimulus inside the RF (ATTin condition) of the neuron compared to when the monkey attends to a stimulus outside the RF (ATTout condition). The bottom shows responses in the thalamic reticular nucleus (TRN), which responds in a reciprocal manner to the LGNm neuron exhibiting an enhanced response when attention is allocated to a stimulus outside the RF. Therefore, the TRN might gate responses in the LGN. Reproduced with permission from [18].

the response to a visual stimulus. In a later phase, the TRN response remained unchanged, but attention further enhanced responses in LGN. These results suggest that (i) the TRN serves as the initiator of modulation in the LGN and (ii) attentional modulation begins at an early stage in the LGN. The TRN therefore plays a crucial role in modulating visual signals at a very early stage of processing.

The pulvinar is a hyperconnected nucleus of the thalamus that has been implicated in the function of visual attention based on anatomical [17,61–63], physiological [64], lesion [65–68] and computational [69] studies. It has been shown that a monkey's ability to suppress distractors is diminished when the pulvinar is pharmacologically inactivated via administration of muscimol [64]. Relay neurons have also been identified in the pulvinar by microstimulating the SC and area MT while simultaneously recording from cells in the pulvinar [70]. This study adds to evidence of a subcortical route for visual signals to reach higher cortex via the pulvinar. At the same time, its bidirectional connections with higher cortical areas make it a potentially important structure in mediating TD signals. However, the pulvinar remains an understudied nucleus, and further studies on this particular brain nucleus are warranted.

Subcortical structures, therefore, both modulate signals in areas encoding BU and TD information, such as the LIP and FEF, and receive TD information from higher cortical areas, directly or possibly through the pulvinar. The SC itself is believed to host a priority map, but this priority map might have closer correspondence to representations needed for motor decisions, including occulomotor behavior and head movements. Thalamic nuclei, including the LGN and TRN, modulate visual signals early on, before they reach cortex, and the pulvinar might be a key relay in communicating attentional signals from one region to another. Subcortical structures are also heavily involved and influenced by reward and emotion, as discussed in the following section.

One emerging theme is that disparate modes of processing might exist in the different brain regions identified above. Areas such as the LIP and FEF, and subcortical structures such as the SC, might normally operate in a feature-agnostic mode, encoding salience and facilitating or inhibiting regions of visual space according to behavioral goals, but without regard to detailed visual features. (This does not, however, preclude these areas from developing feature selectivity through operant training [16], conditioning [71] or task demands [72]). Conversely, visual cortices (areas V1, V2, V4), the IT and the PFC might be operating in a feature-committed mode, modulating responses depending on the exact visual features that give rise to BU salience and/or TD relevance. The pulvinar might then serve as a bidirectional translator, converting fine-grained, feature-committed TD signals to coarser, feature-agnostic TD signals and vice versa. This dichotomy between feature-agnostic and feature-committed TD signals gives rise to interesting hypotheses about possible mechanisms in which TD attention exerts its influence on neural responses in sensory cortex, and thus affects attentional allocation and gaze behavior.

## The role of reward and emotion in TD attention

Until recently, studies of visual attention have traditionally tended to avoid non-visual aspects of cortical and subcortical neuronal responses to manipulations of attention. This has begun to change with a small number of psychophysical and electrophysiological studies that have explored the interplay between reward and attention.

To investigate the role of reward in modulating attention-related responses in the LIP, stimulus selection has been dissociated from motor selection in monkeys [71]. With training, LIP neurons exhibit a strong sustained bias toward the location of a conditioned stimulus, even when a saccade in the opposite direction was required to reveal the outcome of the trial. This suggests that LIP neurons encode 'the value of information' [23] and prioritize spatial locations based on this value.

Studies using operant conditioning paradigms demonstrate effects related to improvements in the volitional TD process. However, learning is also the primary method for augmenting the mandatory TD process. It has been shown that the FEF develops systematic biases, akin to a mandatory TD signal, thereby facilitating shifts of attention in the direction of the feature when it is present at any location [73,74]. More recently, a similar tendency was found in humans performing a visual search task in which the target changed on every trial, which therefore precluded subjects from simply learning a limited set of target features [75]. Subjects' performance improved, demonstrating an improved ability to quickly extract information from a brief preview of the target before each trial, and to then use this information to shape TD signals and guide attention. Learning and reward paradigms can therefore influence ability to both generate TD biasing signals (i.e. volitional TD process) and introduce systematic biases (i.e. mandatory TD process).

Reward plays an important role in modulating attentional signals, and the basal ganglia, which consist of dopaminergic nuclei in the substantia nigra pars reticulata (SNr), the caudate and the putamen, are essential in encoding reward signals [76]. The basal ganglia are integrally connected to the occulomotor system through the connection of the SC to the SNr [77]. Reward signals (TD) from frontal cortices are transmitted to the caudate, which then inhibits the SNr, which in turn pauses the tonic inhibition from the SNr to the SC, releasing it from inhibition and enabling saccades [78]. This follows a more general scheme in the CNS in which the basal ganglia circuit continually inhibits movement of all limbs until an explicit command to make a motor movement is received from cortical or subcortical regions. Furthermore, it is also possible that reward plays a strong role in influencing a subcortical salience map that can cause instant occulomotor reflexes.

A recent study has shed new light on the SNr to SC connection by demonstrating that SNr fibers connect not only to excitatory neurons in the SC, but also to local GABAergic neurons in the intermediate layers of the SC [79]. Therefore, the SNr is involved in shaping the balance of inhibition and excitation in the local SC circuit. SC involvement in attentional selection and the strong role of the SNr in reward render the SNr–SC connection an important one because in most studies, especially physiological studies in monkeys, paradigms are based on the elements of operant conditioning and reinforcement learning with a crucial role for reward (see [78,80] for more detailed discussions).

Sensory processing is also amenable to modulation by brain regions encoding emotions. In particular, it is known that the amygdala has reciprocal connections with both early and late visual areas and can thus give priority, through modulation, to stimuli of ecological relevance [81]. Using a combination of functional magnetic resonance imaging (fMRI) and a study of lesion patients, it was found that visual areas such as the fusiform gyrus receive input from the amygdala and exhibit enhanced responses to affective stimuli [82]. Such modulation by emotion matches response enhancement observed through attentional allocation. Furthermore, it has been shown that emotional and attentional modulations can act independently, as observed in patients with lesions of the amygdala, whose fusiform cortex exhibited responses modulated by attention but not emotion [82]. Affective stimuli can therefore impinge on sensory signals independently of attention; however, the very enhancement due to emotional valence might render the stimuli salient and thus draw more attention. Attention and emotion might thus act independently on the sensory signals and the behavioral relevance of these sensory inputs might be determined by the cumulative effects of both attention and emotion.

One proposal for neural mechanisms and regions involved in fusion of affective inputs with purely visual aspects driving attention has recently been suggested based on a search task in human subjects using fMRI [83]. The frontoparietal spatial attention network, consisting of the superior parietal lobule (SPL), the inferior parietal lobule (IPL) and the FEF, was activated when the cue was purely spatial. However, when the cue contained both spatial and emotional information, limbic and subcortical structures including the posterior cingulate

cortex (PCC), the amygdala and the orbitofrontal cortex were activated, in addition to the frontoparietal network. This study also found selectivity in the PCC for responding only to cues that had emotional valence [83]. These results suggest that the cingulate gyrus, which receives inputs from the amygdala and sends outputs to the frontoparietal network, might serve as the gateway for affective inputs to fuse with spatial biasing signals. This gives rise to a TD salience map in the frontoparietal network, complete with affective and spatial priority information.

Although evidence remains limited, a number of studies have demonstrated links between the attentional network and reward and emotional centers. Such connections must be taken into account when considering TD networks, because most experimental paradigms involving TD attention to date have used reward and/or emotional valence to train and motivate human or animal participants.

## The role of oscillatory activity and neuromodulation in TD attention

It has recently been suggested that synchronous activity (in the gamma range, 50–80 Hz) between cortical regions might serve as the basis for attentional facilitation and cortical computations [84]. In this proposal, neuronal populations representing inputs and decision centers all consist of rhythmically active neural ensembles with distinct excitatory and inhibitory phases. Inhibitory interneurons in each ensemble rhythmically inhibit excitatory pyramidal neurons, thereby establishing a rhythm. Two neural ensembles can then synchronize through phase-locking. This gives rise to a winner-take-all mechanism among two competing inputs feeding into a single higher cortical decision area, through synchronization between the higher area and one selected input. Synchrony between the input and higher areas can be established in a TD or BU manner. In the TD case, a region in higher cortical regions might establish a gamma synchrony with a lower sensory area by phase-locking.

Data from several studies demonstrate that gamma oscillations in the cortex are correlated with attention [45,85]. Disparate brain regions might synchronize their activity in the gamma band when an animal is attending to a particular stimulus. A specific example of this type of coupling is that observed between the FEF and area V4 in monkeys. When attending to a stimulus, coupling through gamma oscillations during attention was observed between neurons in the FEF and V4 [45]. Oscillations in lower frequency bands, such as the alpha and delta bands, have also been implicated in sensory selection [86]. Specifically, in the presence of rhythmic stimuli, delta band oscillations in visual cortex entrain to the rhythm of the stimuli [86]. In doing so, periods of excitability in sensory cortex are aligned with events in the attended stream. In this manner, behaviorally relevant events in the input can be detected more reliably. The same study also showed that the phase of the low-frequency band can modulate amplitudes in higher-frequency bands, such as the gamma band essential for attention. Thus, oscillations in both the gamma and lower-frequency bands are essential neural mechanisms for sensory selection and attention.

The neurochemical basis for attention further supports the notion that synchrony is a possible mechanism for TD attention. Several studies have described acetylcholine (ACh) as the major neurotransmitter involved in mediating attention at the neuronal level [87]. Using pharmacological manipulations, it was found that attentional modulation in area V1 could be enhanced by low doses of ACh [88]. Furthermore, injection of a muscarinic ACh receptor (mAchR) antagonist eliminated such facilitation, but a nicotinic ACh receptor (nAChR) antagonist did not. This demonstrates that ACh acts through mAChRs to modulate attention. Such modulation might enhance processing in sensory areas, a property of TD attention. It has been demonstrated that pharmacological modulation of glutamatergic transmission in the PFC causes an increase in cholinergic release in the PPC [89]. Given the evidence from the studies discussed above, it is reasonable to hypothesize that one neurochemical process by which the PFC could be involved in TD biasing is modulation of ACh release in sensory areas.

One method that has been suggested for achieving gamma synchrony is the disinhibition of pyramidal cells from inhibitory interneuron activity through cholinergic inputs [90]. This suggests that the cholinergic system might also give rise to the gamma synchrony correlated with attention [84]. Taken together, this evidence suggests that one possible mechanism involved in the selection of relevant sensory stimuli is via modulation of ACh by higher cortical regions, such as the PFC, onto sensory cortical regions, which in turn would induce more powerful gamma synchronies between sensory and higher cortical regions. However, it is currently unclear whether gamma synchrony modulation or firing rate modulation is the core mechanism involved in TD attention. This question was addressed using a biophysically realistic computational model of a single layer of visual cortex receiving attentional inputs [91]. The model of the visual cortex consisted of neurons with glutamatergic synapses. These synapses were modeled with two types of glutamate receptors, AMPA and NMDA. Modulation of the ratio of AMPA to NMDA receptor conductance gave rise to both firing rate and gamma synchrony modulation in an independent manner. This suggests that TD attention might be able to regulate these two systems in an independent manner to set or modify gain in sensory areas. Despite the paucity of conclusive empirical evidence, neural gamma synchrony and the concept of glutamatergic modulation in PFC giving rise to ACh modulation in sensory areas provide a compelling potential neural mechanism for TD attention.

## Computational modeling

Physiological studies have guided several theoretical and computational models of attention. Building on the influential feature integration theory [2], guided search theory hypothesizes that massively parallel pre-attentive processes can be guided by TD biasing for features and locations [92]. This theory brings TD elements to a basic BU model of attention [93], which computes individual features at different scales and then combines these features to form a saliency map. A unifying normalization model of attention has recently been proposed and accounts for many effects of TD attention onto visual areas (Figure 6a) [37]. In this model, the neuronal population response of sensory cortex

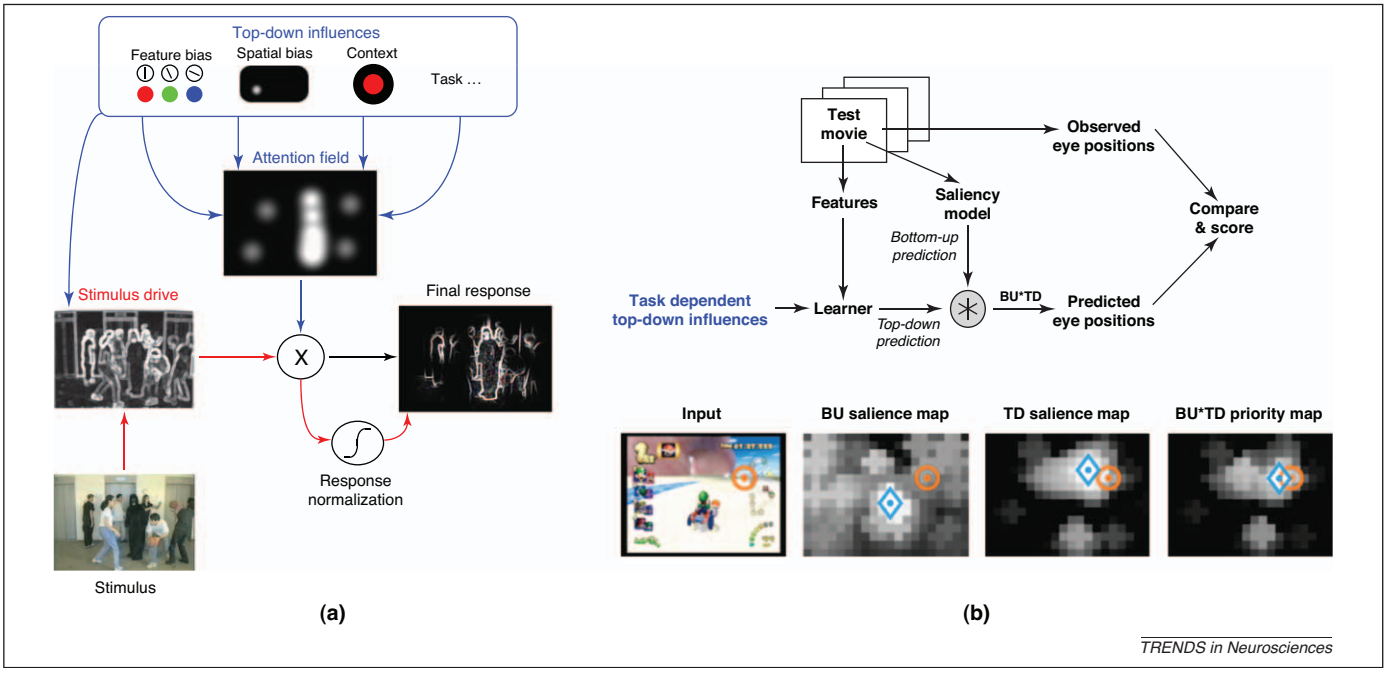



**Figure 6**. Computational modeling of TD attention. **(a)** Model of attention processing inspired by the normalization model of attention [37]. A visual stimulus can be processed by early visual processing stages and this gives rise to stimulus drive. Stimulus drive can then be combined with an attention field that can provide TD modulation over space. Although the model does not specify how the attention field is formed, we hypothesize that TD influences are responsible for this. Note how some TD signals might directly modulate or shape the stimulus drive (e.g. by shifting receptive fields or affecting orientation preference). After combination with the attentional field, responses undergo divisive normalization and contrast gain control before outputting the final response. Figure adapted with permission from [37]. Example input frame provided by Daniel Simons [138]. **(b)** A model based on related ideas that provides a computer implementation applied to the analysis of human gaze behavior while engaged in complex naturalistic tasks (e.g. driving) [97]. A task-dependent learner component builds, during a training phase, associations between distinct coarse types of scenes and observed eye movements (e.g. drivers tend to look to the right when the road turns right). During testing, exposure to similar scenes gives rise to a TD salience map (similar to the attention field in **(a)**), which is further combined with a BU salience map (similar to the stimulus drive) to give rise to the final BU*TD priority map that guides attention. Blue diamonds represent the peak location for each map and orange circles represent the current eye position of a human player. Figure adapted with permission from [97].

to a stimulus is determined by a competitive normalization process that combines stimulus drive, suppressive drive and an attention field. Although this model successfully captures a wide range of single-unit observations, it does not elucidate how the attention field is formed. This concept is related to the idea of a task-relevance map, a topographical map of visual space that might highlight locations or features of current behavioral relevance and might then act as a mask or filter over the BU salience map [94]. The task-relevance map might be populated by combining information about desired features (e.g. look for red items), cued spatial locations (e.g. instructions that the target is to the right), scene gist and context (e.g. when looking for a stapler in an office, focus first on desktops), short-term memory of objects and features at previously visited locations, and TD expectations arising from reasoning about what has been discovered so far in light of the task (e.g. if searching for a computer mouse, finding a keyboard and reasoning that the owner of the machine might be right-handed might bias attention to the right of the keyboard) [94]. Interestingly, recent human neuroimaging data provide direct support for such task relevance or TD salience map possibly located in the intraparietal sulcus (IPS). Indeed, it has been shown that the latter combines, into a single topographic (or, at least, lateralized) map, information about both TD-relevant locations and TD-relevant features [95], and emotional or motivational value of a cued target [83,96]. In a biologically inspired large-scale computer vision implementation, a similar combination of a TD attention field and BU saliency map was used to predict eye movements of humans engaged in complex tasks (e.g. combat flying or first-person exploration video games) [97]. Given the complexity of these tasks and the multiple interacting TD goals involved, this model did not attempt to fully analyze and recognize all objects in scenes and to assess them in light of the task goals. Instead, the TD map was obtained from learned associations between particular types of scenes (summarized by a simple vector of features capturing their gist) and the locations that humans looked at when engaged in the same task and exposed to similar visual scenes (Figure 6b).

At one extreme, TD attention signals might just consist of a single bit of information – to 'enhance' or not – with target visual areas interpreting it in different manners depending on context and on visual inputs. One advantage of such a solution is the low TD communication bandwidth, but an obvious drawback is the inflexibility of signal content. At the other extreme, the brain areas where TD signals originate might address every sensory neuron individually and explicitly modulate the neuron's activity; for example, increasing gain by some specific amount, sharpening tuning, and increasing baseline activity. Such a scheme would afford maximal flexibility, but at the cost of both enormous TD communication bandwidth and high computational requirements in areas where TD signals originate, to compute the exact values for all these signals. The true nature of TD signals is likely to lie between these two extremes, as further elaborated below.

The Guided Search model lies towards the low-bandwidth end of the spectrum, with TD signals imposing spatial attention modulation over coarse regions of visual space and coarse visual features (e.g. a single TD attention weight for each of red, green, blue or yellow colors, or steep, shallow, left or right orientations) [92]. Two recent studies have refined this proposal. First, in human eye-tracking experiments it has been shown that attention and gaze can effectively be guided towards rather fine sub-bands of basic visual features, such as mid-luminance items among low- and high-luminance items, and similarly for size and color saturation [98]. Furthermore, these results have been formalized with a signal-to-noise ratio (SNR)-maximizing model for feature search, whereby the TD gain applied to each sensory neuron is proportional to its ability to distinguish the target of behavioral interest from background clutter [99]. Taken together, these two studies suggest that the bandwidth or granularity of TD signals is unlikely to be extremely low, but rather might consist of at least a few bits for each fine-grained feature sub-band, sufficient to convey optimal biases from the top down. The bandwidth (and number of descending connections) might be higher if different biases can be communicated to different locations of sensory space. At the high extreme, the aforementioned normalization model of attention assumes a highly detailed attention field over space and features [37], implying high-bandwidth TD signals.

Beyond the nature and bandwidth of information conveyed from the top down, computational models have proposed a number of connectivity styles that might be embodied in the biological reality of TD connections. On the one hand, one model has identified a specific dedicated structure (the pulvinar) as a hub or relay for TD signals to reach target visual areas [69]. On the other hand, a more distributed model suggests that TD signals are embedded within the visual areas themselves [100]. In this model, a stimulus is selected at the top level based on an initial sweep of feed-forward information. The spatial selection signals then propagate back and tune lower levels of the (cortical) visual processing hierarchy through a cascade of winner-take-all mechanisms. This view involves retrograde propagation of signals over the processing hierarchy as opposed to direct connections (or through one or a few relays) between top and bottom. A number of models also give specific roles to direct or indirect connections among different levels of the hierarchy, for example between the PFC, FEF, TE and V4 [101]. These models are important because they develop hypotheses for the meaning of large-scale connectivity between brain areas, and these are beginning to be explicitly tested in biological networks using graph-theoretic analyses [102]. Nevertheless, there is a clear lack of specific computational (and experimental) studies that systematically investigate the granularity, bandwidth and specific wiring of TD signals.

Finally, computational theories and models have started to provide hypotheses for the meaning of TD signals. For example, models based on feedback connections from higher cortical areas have been placed in a Bayesian framework, with the suggestion of a generative model that produces a hypothesis about a percept (the prior), then combines this with evidence from BU information to make a final decision on the percept [103]. This approach has been formalized in a hierarchical Bayesian framework [104]. Although these ideas have so far been explored more in the context of the mandatory TD process, they can also

**Box 1. Outstanding questions**

- What is the bandwidth of the TD signal transmitted from one region of the brain to the next? Figure I illustrates the two types of signal. A narrow-bandwidth signal (yellow arrow) defines single weights for individual features, whereas a broadband signal (blue arrow) defines the distribution of gain and tuning over the feature space, as well as the interactions within a feature dimension.
- Are TD signals relayed to visual areas through a central hub (e.g. the pulvinar) or does a more distributed mechanism reflect the reality of communication of TD signals to sensory areas?
- What is the representation or encoding of TD signals? In concrete terms, how are behavioral goals represented and communicated to sensory neurons that are tuned to specific features?
- What (if any) computations take place subcortically, independent of the cortex, that would influence attention modulation of sensory perception?

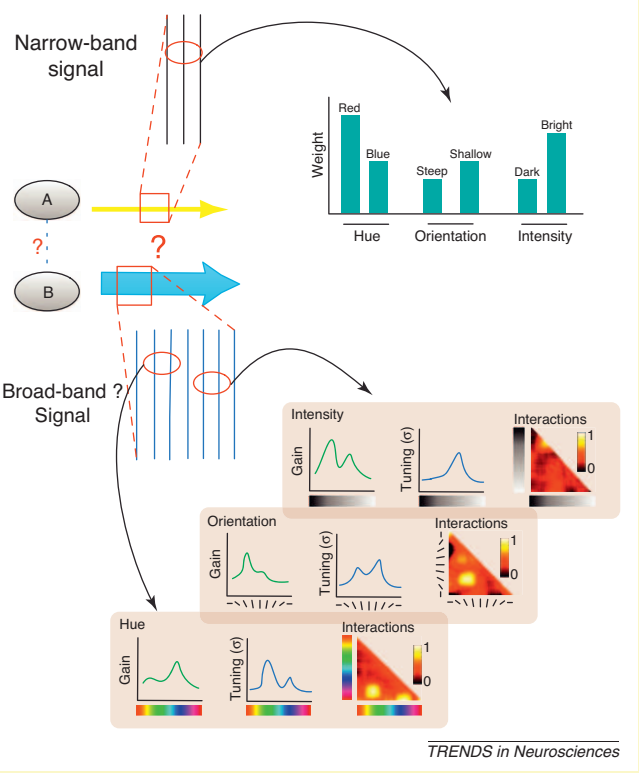



**Figure I.** Narrow-band versus broad-band TD biasing signals. The TD biasing signals transmitted from one area 'A' of the brain to another area 'B' can be either narrow-band (yellow arrow) or broad-band (blue arrow) in nature. Narrow-band signals consist of a small set of weights that bias feature preferences in a coarse manner. The bar graph shows a signal that applies a higher gain to neurons tuned to red rather than blue in the color feature dimension, neurons tuned to shallow rather than steep orientations, and neurons tuned to brighter rather than darker stimuli. Broad-band biasing signals (bottom) contain a greater amount of information and might facilitate biasing of features in a detailed manner, weighing gain, tuning and feature interactions independently. Rather than simply setting a weight along a feature dimension, as is the case in the narrow-band example, broad-band TD signals might set a biasing profile along the feature dimension, as shown in the example graphs. Green curves show a biasing profile for gains of neurons along a feature dimension. Blue curves show a biasing profile of tuning of neurons; a peak here would indicate a bias or shift of tuning of neurons for the particular feature value. The interaction triangles on the right show biases for feature interactions. For example, along the hue dimension, there are two hot spots, one indicating a preference for simultaneous occurrence of yellow and red hues and another indicating a preference for red and blue hues.

be placed in the context of the volitional TD process. Volitional TD control could then be understood as updating, biasing or disambiguating the prior based on high-level tasks, contextual cues or behavioral goals. In computer-vision models using these principles, it has indeed been shown that TD attention provides great benefit over pure BU processing [105]. For example, TD information can more effectively guide visual search for specific objects in natural scenes (e.g. pedestrians in street scenes) by limiting the search to spatial locations of high prior or posterior probabilities [106–108]. Although computational models have made some headway in both incorporating experimental data and generating predictions to guide further experiments, much remains to be done both experimentally and theoretically to unravel the mechanisms by which TD attentional mechanisms influence BU processing (Box 1).

## Conclusion

Attention modulates sensory signals early in the process, exerting its influence on the SC and the thalamus before further modulating signals in cortex. The cumulative effects of this modulation based on both TD and BU influences might be represented by a priority map over visual space. Although there is some debate about the exact locus of the priority map, it is clear that the LIP, FEF and SC exhibit properties that are compatible with the existence of a spatial map encoding behavioral relevance of spatial locations. These three regions might jointly compute or host such a map that is agnostic to the features that caused the priority. Thus, the map fuses both BU and TD influences and drives motor output.

Higher cortical areas such as the PFC send detailed TD signals to sensory areas for biasing of spatial and non-spatial features. Such signals fuse together with reward-related and emotional signals to form the TD influence on attention, which might be reflected in the priority map. Subcortical regions, through their close connection to the reward systems in the brain and their coupling with motor systems, exert strong influences on attentional signals, in addition to being major targets of attentional modulation for motor output. Feedback connections are both pervasive and crucial for the transmission of biasing signals emanating from higher brain regions, especially the frontal cortices that are involved in working memory processes and send descending reward signals. Computational studies highlight the important constraints on the nature, granularity, bandwidth and connectivity style of TD connections. There is a pressing need to build models that take into account physiological data, particularly from microstimulation and lesion studies, which could help to determine the contributions of specific areas to the computations necessary for attentional guidance.

Although the exact mechanisms of TD attention have yet to be completely delineated, there are sufficient data available to demonstrate that attention is mediated by the merging of TD and BU information. As William James eloquently stated, 'The attentive process, therefore, at its maximum may be physiologically symbolized, by a brain-cell played on in two ways from without and from within' [109].

## Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), General Motors Corporation, and the Army Research Office (grant no. W911NF-08-1-0360). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof. We would also like to thank Robert Desimone, Jack Gallant, Jacqueline Gottlieb and the anonymous reviewers for their helpful comments and suggestions.

## References

1 James, W. (1890) *The Principles of Psychology*, Harvard University Press
2 Treisman, A. and Gelade, G. (1980) A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136
3 Wolfe, J.M. and Horowitz, T.S. (2004) What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501
4 Itti, L. and Koch, C. (2001) Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203
5 Gilbert, C. and Sigman, M. (2007) Brain states: top-down influences in sensory processing. *Neuron* 54, 677–696
6 Chun, M.M. and Jiang, Y. (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cogn. Psychol.* 36, 28–71
7 Desimone, R. and Duncan, J. (1995) Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222
8 Noudoost, B. *et al.* (2010) Top-down control of visual attention. *Curr. Opin. Neurobiol.* 20, 183–190
9 Miller, E.K. and Cohen, J.D. (2001) An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202
10 Corbetta, M. and Shulman, G.L. (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215
11 Werblin, F. *et al.* (2001) Parallel processing in the mammalian retina: lateral and vertical interactions across stacked representations. *Prog. Brain Res.* 131, 229–238
12 Goodale, M.A. and Milner, A.D. (1992) Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25
13 Wurtz, R. and Mohler, C. (1976) Enhancement of visual responses in monkey striate cortex and frontal eye fields. *J. Neurophysiol.* 39, 766–772
14 Fecteau, J. and Munoz, D. (2006) Salience, relevance, and firing: a priority map for target selection. *Trends Cogn. Sci.* 10, 382–390
15 Bisley, J.W. and Goldberg, M.E. (2010) Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.* 33, 1–21
16 Thompson, K.G. and Bichot, N.P. (2005) A visual salience map in the primate frontal eye field. *Prog. Brain Res.* 147, 251–262
17 Shipp, S. (2003) The functional logic of cortico-pulvinar connections. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 358, 1605–1624
18 McAlonan, K. *et al.* (2008) Guarding the gateway to cortex with attention in visual thalamus. *Nature* 456, 391–394
19 O'Connor, D. *et al.* (2002) Attention modulates responses in the human lateral geniculate nucleus. *Nat. Neurosci.* 5, 1203–1209
20 Buffalo, E.A. *et al.* (2010) A backward progression of attentional effects in the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* 107, 361–365
21 McAdams, C.J. and Maunsell, J.H. (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.* 19, 431–441
22 Martinez-Trujillo, J.C. and Treue, S. (2004) Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr. Biol.* 14, 744–751
23 Gottlieb, J. and Balan, P. (2010) Attention as a decision in information space. *Trends Cogn. Sci.* 14, 240–248
24 Gottlieb, J. (2007) From thought to action: the parietal cortex as a bridge between perception, action, and cognition. *Neuron* 53, 9–16
25 Gottlieb, J. *et al.* (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484
26 Bisley, J. and Goldberg, M. (2003) Neuronal activity in the lateral intraparietal area and spatial attention. *Science* 299, 81–86
27 Ungerleider, L.G. *et al.* (2008) Cortical connections of area V4 in the macaque. *Cereb. Cortex* 18, 477–499

28 Lamme, V.A. and Roelfsema, P.R. (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579
29 Monosov, I. *et al.* (2008) Measurements of simultaneously recorded spiking activity and local field potentials suggest that spatial selection emerges in the frontal eye field. *Neuron* 57, 614–625
30 Buschman, T. and Miller, E. (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315, 1860–1862
31 Lebedev, M.A. *et al.* (2004) Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biol.* 2, e365
32 Luck, S.J. *et al.* (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42
33 Bichot, N.P. *et al.* (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308, 529–534
34 Saenz, M. *et al.* (2003) Global feature-based attention for motion and color. *Vis. Res.* 43, 629–637
35 Treue, S. and Martinez Trujillo, J.C. (1999) Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* 399, 575–579
36 Motter, B.C. (1994) Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* 14, 2178–2189
37 Reynolds, J.H. and Heeger, D.J. (2009) The normalization model of attention. *Neuron* 61, 168–185
38 Reynolds, J. and Chelazzi, L. (2004) Attentional modulation of visual processing. *Annu. Rev. Neurosci.* 27, 611–647
39 Knudsen, E.I. (2007) Fundamental components of attention. *Annu. Rev. Neurosci.* 30, 57–78
40 Tomita, H. *et al.* (1999) Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* 401, 699–703
41 Rossi, A. *et al.* (2007) Top down attentional deficits in macaques with lesions of lateral prefrontal cortex. *J. Neurosci.* 27, 11306–11314
42 Opris, I. *et al.* (2005) Microstimulation of the dorsolateral prefrontal cortex biases saccade target selection. *J. Cogn. Neurosci.* 17, 893–904
43 Moore, T. and Armstrong, K. (2003) Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421, 370–373
44 Winkowski, D. and Knudsen, E. (2008) Distinct mechanisms for top-down control of neural gain and sensitivity in the owl optic tectum. *Neuron* 60, 698–708
45 Gregoriou, G.G. *et al.* (2009) High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science* 324, 1207–1210
46 Wardak, C. *et al.* (2006) Contribution of the monkey frontal eye field to covert visual attention. *J. Neurosci.* 26, 4228–4235
47 Cutrell, E. and Marrocco, R. (2002) Electrical microstimulation of primate posterior parietal cortex initiates orienting and alerting components of covert attention. *Exp. Brain Res.* 144, 103–113
48 Balan, P.F. and Gottlieb, J. (2009) Functional significance of nonspatial information in monkey lateral intraparietal area. *J. Neurosci.* 29, 8166–8176
49 Friedman-Hill, S. *et al.* (2003) Posterior parietal cortex and the filtering of distractors. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4263–4268
50 Saalmann, Y. *et al.* (2007) Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science* 316, 1612
51 Liu, Y. *et al.* (2010) Intention and attention: different functional roles for LIPD and LIPV. *Nat. Neurosci.* 13, 495–500
52 Sherman, S. (2007) The thalamus is more than just a relay. *Curr. Opin. Neurobiol.* 17, 417–422
53 Kawasaki, K. and Sheinberg, D.L. (2008) Learning to recognize visual objects with microstimulation in inferior temporal cortex. *J. Neurophysiol.* 100, 197–211
54 Afraz, S-R. *et al.* (2006) Microstimulation of inferotemporal cortex influences face categorization. *Nature* 442, 692–695
55 Cavanaugh, J. and Wurtz, R. (2004) Subcortical modulation of attention counters change blindness. *J. Neurosci.* 24, 11236–11243
56 Muller, J. *et al.* (2005) Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 524–529
57 Carello, C. and Krauzlis, R., (2004) Manipulating intent evidence for a causal role of the superior colliculus in target selection. *Neuron* 43, 575–583
58 McPeek, R. and Keller, E. (2004) Deficits in saccade target selection after inactivation of superior colliculus. *Nat. Neurosci.* 7, 757–763
59 Ignashchenkova, A. *et al.* (2003) Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. *Nat. Neurosci.* 7, 56–64
60 Lovejoy, L. and Krauzlis, R. (2009) Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. *Nat. Neurosci.* 13, 261–266
61 Kaas, J.H. and Lyon, D.C. (2007) Pulvinar contributions to the dorsal and ventral streams of visual processing in primates. *Brain Res. Rev.* 55, 285–296
62 Leh, S. *et al.* (2008) The connectivity of the human pulvinar: a diffusion tensor imaging tractography study. *Int. J. Biomed. Imaging* 2008, 1–5
63 Robinson, D.L. and Petersen, S.E. (1992) The pulvinar and visual salience. *Trends Neurosci.* 15, 127–132
64 Desimone, R. *et al.* (1990) Attentional control of visual perception: cortical and subcortical mechanisms. *Cold Spring Harb. Symp. Quant. Biol.* 55, 963–971
65 Snow, J. *et al.* (2009) Impaired attentional selection following lesions to human pulvinar: evidence for homology between human and monkey. *Proc. Natl. Acad. Sci. U.S.A.* 106, 4054–4059
66 Arend, I. *et al.* (2008) Spatial and temporal deficits are regionally dissociable in patients with pulvinar lesions. *Brain* 131, 2140–2152
67 Bender, D.B. and Baizer, J.S. (1990) Saccadic eye movements following kainic acid lesions of the pulvinar in monkeys. *Exp. Brain Res.* 79, 467–478
68 Bender, D.B. and Butter, C.M. (1987) Comparison of the effects of superior colliculus and pulvinar lesions on visual search and tachistoscopic pattern discrimination in monkeys. *Exp. Brain Res.* 69, 140–154
69 Olshausen, B.A. *et al.* (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719
70 Berman, R.A. and Wurtz, R.H. (2010) Functional identification of a pulvinar path from superior colliculus to cortical area MT. *J. Neurosci.* 30, 6342–6354
71 Peck, C.J. *et al.* (2009) Reward modulates attention independently of action value in posterior parietal cortex. *J. Neurosci.* 29, 11182–11191
72 Toth, L.J. and Assad, J.A. (2002) Dynamic coding of behaviourally relevant stimuli in parietal cortex. *Nature* 415, 165–168
73 Bichot, N.P. and Schall, J.D. (1999) Effects of similarity and history on neural mechanisms of visual selection. *Nat. Neurosci.* 2, 549–554
74 Bichot, N.P. *et al.* (1996) Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature* 381, 697–699
75 Baluch, F. and Itti, L. (2010) Training top-down attention improves performance on a triple-conjunction search task. *PLoS ONE* 5, e9127
76 Graybiel, A. (2005) The basal ganglia: learning new tricks and loving it. *Curr. Opin. Neurobiol.* 15, 638–644
77 Boehnke, S.E. and Munoz, D.P. (2008) On the importance of the transient visual response in the superior colliculus. *Curr. Opin. Neurobiol.* 18, 544–551
78 Hikosaka, O. *et al.* (2006) Basal ganglia orient eyes to reward. *J. Neurophysiol.* 95, 567–584
79 Kaneda, K. *et al.* (2008) Nigral inhibition of GABAergic neurons in mouse superior colliculus. *J. Neurosci.* 28, 11071–11078
80 Shires, J. *et al.* (2010) Shedding new light on the role of the basal ganglia-superior colliculus pathway in eye movements. *Curr. Opin. Neurobiol.* 20, 717–725
81 Pessoa, L. (2008) On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* 9, 148–158
82 Vuilleumier, P. *et al.* (2004) Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nat. Neurosci.* 7, 1271–1278
83 Mohanty, A. *et al.* (2009) Search for a threatening target triggers limbic guidance of spatial attention. *J. Neurosci.* 29, 10563–10572
84 Fries, P. (2009) Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu. Rev. Neurosci.* 32, 209–224
85 Fries, P. *et al.* (2002) Oscillatory neuronal synchronization in primary visual cortex as a correlate of stimulus selection. *J. Neurosci.* 22, 3739–3754

86 Lakatos, P. *et al.* (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113
87 Sarter, M. *et al.* (2005) Unraveling the attentional functions of cortical cholinergic inputs: interactions between signal-driven and cognitive modulation of signal detection. *Brain Res. Rev.* 48, 98–111
88 Herrero, J.L. *et al.* (2008) Acetylcholine contributes through muscarinic receptors to attentional modulation in V1. *Nature* 454, 1110–1114
89 Nelson, C.L. *et al.* (2005) Prefrontal cortical modulation of acetylcholine release in posterior parietal cortex. *Neuroscience* 132, 347–359
90 Deco, G. and Thiele, A. (2009) Attention: oscillations and neuropharmacology. *Eur. J. Neurosci.* 30, 347–354
91 Buehlmann, A. and Deco, G. (2008) The neuronal basis of attention: rate versus synchronization modulation. *J. Neurosci.* 28, 7679–7686
92 Wolfe, J. (1994) Guided search 2.0. A revised model of visual search. *Psychonom. Bull. Rev.* 1, 202–238
93 Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227
94 Navalpakkam, V. and Itti, L. (2005) Modeling the influence of task on attention. *Vis. Res.* 45, 205–231
95 Egner, T. *et al.* (2008) Neural integration of top-down spatial and feature-based information in visual search. *J. Neurosci.* 28, 6141–6151
96 Mohanty, A. *et al.* (2008) The spatial attention network interacts with limbic and monoaminergic systems to modulate motivation-induced attention shifts. *Cereb. Cortex* 18, 2604–2613
97 Peters, R. and Itti, L. (2007) Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8
98 Navalpakkam, V. and Itti, L. (2006) Top-down attention selection is fine-grained. *J. Vis.* 6, 1180–1193
99 Navalpakkam, V. and Itti, L. (2007) Search goal tunes visual features optimally. *Neuron* 53, 605–617
100 Tsotsos, J.K. *et al.* (1995) Modeling visual-attention via selective tuning. *Artif. Intell.* 78, 507–545
101 Hamker, F.H. (2006) Modeling feature-based attention as an active top-down inference process. *Biosystems* 86, 91–99
102 Bullmore, E. and Sporns, O. (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198
103 Gregory, R. (1970) *The Intelligent Eye*, McGraw-Hill
104 Lee, T. and Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20, 1434–1448
105 Frintrop, S. *et al.* (2010) Computational visual attention systems and their cognitive foundation: a survey. *ACM Trans. Appl. Percept.* 7, 6
106 Najemnik, J. and Geisler, W.S. (2005) Optimal eye movement strategies in visual search. *Nature* 434, 387–391
107 Torralba, A. *et al.* (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113, 766–786
108 Ehinger, K.A. *et al.* (2009) Modeling search for people in 900 scenes: a combined source model of eye guidance. *Vis. Cogn.* 17, 945–978
109 James, W. (1892) *Talks to Teachers on Psychology and to Students on Some of Life's Ideals*, Adamant Media Corporation
110 Rafal, R.D. and Posner, M.I. (1987) Deficits in human visual spatial attention following thalamic lesions. *Proc. Natl. Acad. Sci. U.S.A.* 84, 7349–7353
111 Pezaris, J.S. and Reid, R.C. (2007) Demonstration of artificial visual percepts generated through thalamic microstimulation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7670–7675
112 Schmid, M.C. *et al.* (2010) Blindsight depends on the lateral geniculate nucleus. *Nature* 466, 373–377
113 Page, W.K. *et al.* (1994) Magnocellular or parvocellular lesions in the lateral geniculate nucleus of monkeys cause minor deficits of smooth pursuit eye movements. *Vis. Res.* 34, 223–239
114 Tehovnik, E.J. *et al.* (2003) Differential effects of laminar stimulation of V1 cortex on target selection by macaque monkeys. *Eur. J. Neurosci.* 16, 751–760
115 Tehovnik, E.J. *et al.* (2003) Saccadic eye movements evoked by microstimulation of striate cortex. *Eur. J. Neurosci.* 17, 870–878
116 Murphey, D.K. and Maunsell, J.H.R. (2007) Behavioral detection of electrical microstimulation in different cortical visual areas. *Curr. Biol.* 17, 862–867
117 Moore, T. *et al.* (2001) Direction of motion discrimination after early lesions of striate cortex (V1) of the macaque monkey. *Proc. Natl. Acad. Sci. U.S.A.* 98, 325–330
118 Yoshida, M. *et al.* (2008) Striate cortical lesions affect deliberate decision and control of saccade: implication for blindsight. *J. Neurosci.* 28, 10517–10530
119 Buffalo, E. *et al.* (2005) Impaired filtering of distracter stimuli by TE neurons following V4 and TEO lesions in macaques. *Cereb. Cortex* 15, 141–151
120 Bachevalier, J. and Mishkin, M. (1994) Effects of selective neonatal temporal lobe lesions on visual recognition memory in rhesus monkeys. *J. Neurosci.* 14, 2128–2139
121 Petrides, M. (2000) Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. *J. Neurosci.* 20, 7496–7503
122 Nichols, M.J. and Newsome, W.T. (2002) Middle temporal visual area microstimulation influences veridical judgments of motion direction. *J. Neurosci.* 22, 9530–9540
123 Bisley, J.W. *et al.* (2001) Microstimulation of cortical area MT affects performance on a visual working memory task. *J. Neurophysiol.* 85, 187–196
124 Newsome, W.T. and Paré, E.B. (1988) A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* 8, 2201–2211
125 Hanks, T.D. *et al.* (2006) Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nat. Neurosci.* 9, 682–689
126 Moore, T. and Fallah, M. (2004) Microstimulation of the frontal eye field and its effects on covert spatial attention. *J. Neurophysiol.* 91, 152–162
127 Juan, C. *et al.* (2004) Dissociation of spatial attention and saccade preparation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15541–15544
128 Ruff, C.C. *et al.* (2006) Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. *Curr. Biol.* 16, 1479–1488
129 Neggers, S.F.W. *et al.* (2007) TMS pulses on the frontal eye fields break coupling between visuospatial attention and eye movements. *J. Neurophysiol.* 98, 2765–2778
130 Wegener, S.P. *et al.* (2008) Microstimulation of monkey dorsolateral prefrontal cortex impairs antisaccade performance. *Exp. Brain Res.* 190, 463–473
131 Rasmusson, D.D. *et al.* (2007) Inactivation of prefrontal cortex abolishes cortical acetylcholine release evoked by sensory or sensory pathway stimulation in the rat. *Neuroscience* 149, 232–241
132 Rudolph, K. and Pasternak, T. (1999) Transient and permanent deficits in motion perception after lesions of cortical areas MT and MST in the macaque monkey. *Cereb. Cortex* 9, 90–100
133 De Weerd, P. *et al.* (1999) Loss of attentional stimulus selection after extrastriate cortical lesions in macaques. *Nat. Neurosci.* 2, 753–758
134 Gregory, R. (1997) Knowledge in perception and illusion. *Philos. Trans. R. Soc. B Biol. Sci.* 352, 1121–1127
135 Van Essen, D.C. (2002) Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr. Opin. Neurobiol.* 12, 574–579
136 Van Essen, D.C. *et al.* (2001) An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Inform. Assoc.* 8, 443–459
137 Pascual-Leone, A. and Walsh, V. (2001) Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science* 292, 510–512
138 Simons, D.J. and Chabris, C.F. (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28, 1059–1074

# Exploiting Local and Global Patch Rarities for Saliency Detection

Ali Borji
USC
borji@usc.edu

Laurent Itti
USC
itti@usc.edu

## Abstract

*We introduce a saliency model based on two key ideas. The first one is considering local and global image patch rarities as two complementary processes. The second one is based on our observation that for different images, one of the RGB and Lab color spaces outperforms the other in saliency detection. We propose a framework that measures patch rarities in each color space and combines them in a final map. For each color channel, first, the input image is partitioned into non-overlapping patches and then each patch is represented by a vector of coefficients that linearly reconstruct it from a learned dictionary of patches from natural scenes. Next, two measures of saliency (Local and Global) are calculated and fused to indicate saliency of each patch. Local saliency is distinctiveness of a patch from its surrounding patches. Global saliency is the inverse of a patch's probability of happening over the entire image. The final saliency map is built by normalizing and fusing local and global saliency maps of all channels from both color systems. Extensive evaluation over four benchmark eye-tracking datasets shows the significant advantage of our approach over 10 state-of-the-art saliency models.*

## 1. Introduction

The human visual system has to process an enormous amount of incoming information ($\sim 10^8$ bit/s) from the retina. Similarly, in computer vision, many systems suffer from the high computational complexity of inputs, especially when these systems are supposed to work in real time. Visual saliency is a concept that offers efficient solutions for both biological and artificial vision systems. It is basically a process that detects scene regions different from their surroundings (often referred as bottom-up saliency). Then, higher cognitive and usually more complex operations are focused only on the selected areas.

Recently, modeling visual saliency has raised much interest in theory and applications (see [47] for a review). For example in computer vision, it has been used for image and video compression [49], image segmentation, and object recognition [52]. In computer graphics, detecting salient regions has been employed for content-aware image cropping, photo collage [50], and stylization of images [53]. Saliency computation has also applications in other areas such as advertisement design [51] and visual prosthetics [48]. Our focus in this paper is proposing a new and more predictive (with respect to human eye tracking data) model of bottom-up visual saliency by integrating local and global saliency detection in both RGB and Lab color spaces (see Fig. 1).

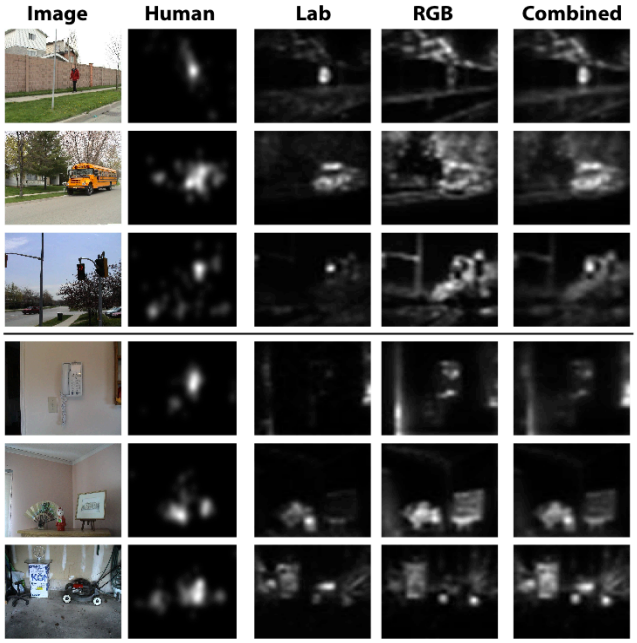



Figure 1. **One color system does not work for all images.** Top (Bottom): Sample images where our model is able to detect the outliers in CIE Lab (RGB) color space. For some images both color spaces work equally well. Last column shows combined maps from both color spaces. Images are taken from the TORONTO dataset [14].

**Related works on saliency modeling**. A majority of computational models of attention follow the structure adapted from the Feature Integration Theory (FIT) [15] and the Guided Search model [1]. Koch and Ullman [19] proposed a computational architecture for this theory and Itti *et al.* [4] were among the first ones to fully implement and maintain it. The main idea here is to compute saliency in

each of several features (e.g., color, intensity, orientation; saliency is then the relative difference between a region and its surrounding) in parallel, and to fuse them in a scalar map called the "saliency map". Le Meur *et al.* [18] adapted the Koch-Ullman's model to include features of contrast sensitivity, perceptual decomposition, visual masking, and center-surround interactions. Some models have added features such as symmetry [20], texture contrast [36], curvedness [21], or motion [41] to the basic structure.

In addition to the mentioned cognitive models, several probabilistic models of visual saliency have been developed over the past years. In these models, a set of statistics or probability distributions are computed from either the current scene, or from a set of natural scenes over space or time or both. Itti and Baldi [10] defined surprising stimuli as those which significantly change beliefs of an observer, measured as the Kullback-Leibler (KL) distance between posterior and prior beliefs. Harel *et al.* [7] used graph algorithms and a measure of dissimilarity to achieve efficient saliency computation with their Graph Based Visual Saliency (GBVS) model. Torralba *et al.*'s contextual guidance model [26] consolidates low-level salience and scene context when guiding search. Areas of high salience within a selected contextual region are given higher weights on an activation map than those that fall outside the selected contextual region. Some Bayesian models formulate visual search and derive a measure of bottom-up saliency as a by-product. For example, Zhang *et al.*'s model [12], Saliency Using Natural statistics (SUN), combines top-down and bottom-up information to guide eye movements during real-world object search tasks. However, unlike Torralba *et al.*'s model, SUN implements target features as the top-down component. Gao and Vasconcelos [23] define saliency as maximizing classification accuracy. They utilize the KL distance to measure mutual information between features at a scene location and class labels. The higher mutual information between a region and class of interest, the higher the saliency of that region. Seo and Milanfar [11] using local regression kernels build a "self-resemblance" map, which measures the similarity of a feature matrix at a pixel of interest to its neighboring feature matrices.

Bruce and Tsotsos [14] proposed the Attention based on Information Maximization (AIM) model by employing the first principles of information theory. They model bottom-up saliency as the maximum information sampled from an image. More specifically, saliency is computed as Shannon's self-information $-log\ p(f)$, where $f$ is a local visual feature. Hou and Zhang [9] introduced the Incremental Coding Length (ICL) approach to measure the respective entropy gain of each feature. The goal is to maximize the entropy of the sampled visual features.

Some models measure saliency in the frequency domain. Hou and Zhang [8] propose a method based on relating extracted spectral residual features of an image in the spectral domain to the spatial domain. Guo *et al.* [24] show that incorporating the Phase spectrum of the Quaternion Fourier Transform (PQFT) instead of the amplitude transform leads to better saliency predictions in the spatio-temporal domain.

Some models learn saliency. Kienzle *et al.* [2] utilize support vector machines (SVM) to learn saliency of each image patch directly from human eye tracking data. Similarly, Judd *et al.* [3] train a linear SVM from human eye movement data, using a set of low, mid, and high-level image features to define salient locations. Feature vectors from highly fixated locations are assigned class label $+1$ while less fixated locations are assigned label $-1$. Zhao and Koch [22] used least-squares regression to learn the weights associated with a set of feature maps from subjects freely fixating natural scenes drawn from four different eye-tracking data sets. They find that the weights can be quite different for different data sets, but their face-detection and orientation channels are usually more important than color and intensity channels. Navalpakkam and Itti [29] define visual saliency in terms of signal to noise ratio (SNR) of a target object versus background and learn parameters of a linear combination of low-level features that cause the highest expected SNR for detecting a target from distractors.

**Contributions.** The models reviewed above fall into two general categories: 1) models that calculate saliency by implementing local center-surround operations (e.g., Itti *et al.* [4], Surprise [10], Judd *et al.* [3], GBVS [7], and Rahtu *et al.* [39]), 2) models that find salient regions globally by calculating rarity of features over the entire scene (e.g., AIM [14], SUN [12], Torralba [26], SRM [8], ICL [9], and Rarity model [25]). Our first contribution is to propose a unified model that benefits from the advantages of both approaches, which thus far have been treated independently. Note that the ideas of local and global context have been (separately) considered in the past [44][17] by salient object detection/segmentation approaches, but those have not yet been tested with human fixation prediction, which is the goal of most models (including ours).

Almost all saliency approaches utilize a color channel. Some have used RGB (e.g., [4][3][14][12]) while others have employed Lab (e.g., [42][18][39]), inspired by the finding that it better approximates human color perception. In particular, Lab aspires to perceptual uniformity, and its L component closely matches human perception of lightness, while the a and b channels approximate the human chromatic opponent system. RGB, on the other hand, is often the default choice for scene representation and storage. We argue that employing just one color system does not always lead to successful outlier detection. In Fig.1, we show that interesting objects in some images are more salient in Lab color space, while, for some others, saliency detection works better in RGB. Hence, a yet unexplored strat-

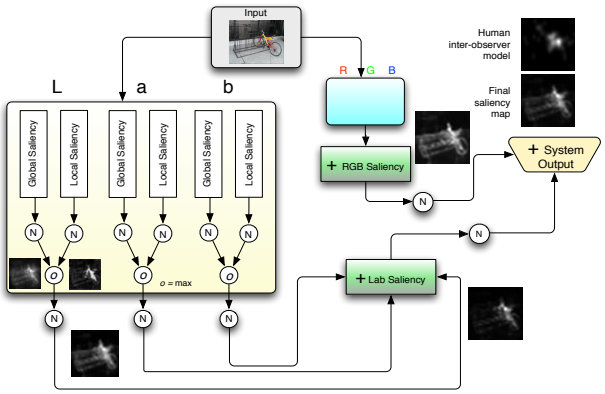



Figure 2. **Diagram of our proposed model.** First, the input image is transformed into Lab and RGB formats. Then, in each channel of a color space, a global saliency map based on rarity of an image patch in the entire scene, and a local saliency map, the dissimilarity between a patch and its surrounding window, are computed, normalized, and combined. Outputs of color channels (i.e., L, a, or b, similarly for RGB) are normalized and combined once more to form the output of a color system. The final map is the summation of the normalized maps in two color spaces.

egy, which is our second contribution, is combining saliency maps from both color spaces.

We compare accuracy of our model and its subcomponents with the mainstream models over four benchmark eye tracking datasets. These are top-ranked models that previous studies have shown to be significantly predictive of eye fixations in free viewing of natural scenes.

## 2. Proposed Saliency Model

Our proposed framework is presented in Fig. 2. An input image in two formats (Lab and RGB) undergoes the same saliency detection and the resultant maps in each color system are normalized and summed. In each color format, two local and global saliency operations are applied to each color sub-channel separately. While the first operation detects outliers in a local surrounding, the latter calculates the rarity of a feature or a region over the entire scene. Then, local and global rarities are combined to generate the output of each channel. Channel output maps are then normalized and summed once more to generate the saliency map. The whole process can be performed over several scales. There is no need to directly calculate the orientation channel in our model (since some patches from the chosen ensemble will emulate it; see below).

There is a large body of behavioral support for both local and global operations from the cognitive science literature. While early studies favored the thesis that local contrast attracts attention [15][19][4], recent work has shifted toward understanding top-down conceptual factors which seem to operate at the object level (see Figs. 1 and 7 for some examples). Such factors in free-viewing include human body, signs, cars, faces and text [3][45]. Particularly, Einhäuser *et al.* [30], showed that objects predict human fixations better than low-level saliency. Also it has been shown that interesting objects are more salient within a scene, providing support in favor of object-based attention [43]. As rare features are more likely to belong to a single object and since objects are rare compared to background in natural scenes, we believe that global saliency can help detecting top-down object-level concepts. Thus, instead of leaning on only one component (local or global), an effective strategy is integrating both of these complementary processes.

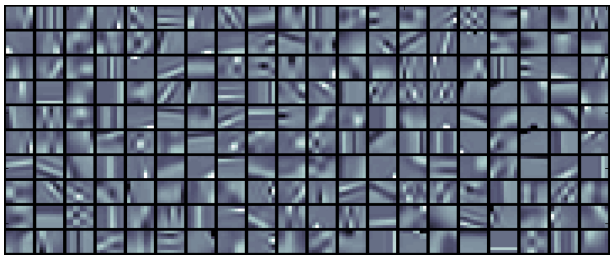


Figure 3. **A dictionary of 200 basis functions** learned from a large repository of natural images for the L channel of the Lab color space. Image size and patch size ($w$) were $512 \times 512$ and $8 \times 8$, respectively.

We estimate saliency on a patch-by-patch basis: each image patch is projected into the space of a dictionary of image patches (basis functions) learned from a repository of natural scenes. Each patch of an image is then represented by a vector of basis coefficients that can linearly reconstruct it.

### 2.1. Image representation

It is well known that natural images can be sparsely represented by a set of localized and oriented filters [27][28]. Also, recent progress in computer vision has demonstrated that sparse coding is an effective tool for image representation for several applications such as image classification [31][32], face recognition [33], image denoising [34], as well as saliency detection [14][12][9]. The underlying idea behind sparse coding is that a vision system should be adapted based on statistics of the visual environment where it is supposed to operate. As a supporting evidence for this theory, it has been shown that receptive fields (RF) of some neurons in V1 cortex resemble those RFs that are learned by sparse coding algorithms [27].

Mathematically, given a set of $n$ $m-$dimensional basis signals (dictionary) $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots \mathbf{d}_n] \in \mathbb{R}^{m \times n}$, the sparse coding of an input signal $\mathbf{x} \in \mathbb{R}^m$ can be found by solving an "$l_1$-norm minimization problem":

$$\boldsymbol{\alpha}^*(\mathbf{x}, \mathbf{D}) = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2} ||\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda_1 ||\boldsymbol{\alpha}||_1 \qquad (1)$$

where $||.||_1$ denotes the $l_1$-norm and $\lambda_1$ is a regularization parameter. Thus, $\mathbf{x} \sim \mathbf{x}^{'} = \mathbf{D}\alpha^*$ where $\mathbf{x}^{'}$ is the estimation of $\mathbf{x}$. To learn the dictionary $\mathbf{D}$, considering a training set of $q$ data samples $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots \mathbf{y}_q]$ in $\mathbb{R}^{m \times q}$ an empirical cost function $g_q(\mathbf{D}) = \frac{1}{q} \sum_{i=1}^{q} l_u(\mathbf{y}_i, \mathbf{D})$ is minimized,

where $l_u(\mathbf{y}, \mathbf{D})$ is:

$$l_u(\mathbf{y}_i, \mathbf{D}) = min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2}||\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}||_2^2 + \lambda_1||\boldsymbol{\alpha}||_1 \quad (2)$$

We represent an image patch by a linear combination of some basis functions which correspond or act as feature detectors in early visual areas of the brain (neuron receptive fields or transfer functions). Given an input image, it is first resized to $2^w \times 2^w$ pixels where patch size $w$ is selected in a way that $2^w$ is divisible to $w$. Let $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots \mathbf{p}_n\}$ represent the set of linearized image patches from top-left to bottom-right with no overlap. Then using Eq. 1, coefficients that reconstruct each patch are calculated and are used to represent that patch. By reshaping reconstructed patches and aligning them, the original image can be reproduced.

To learn a dictionary of patch bases (i.e., minimizing $g_q(\mathbf{D})$), we extracted $500,000$ $8 \times 8$ image patches (for each sub channel of RGB or Lab) from 1500 randomly selected color images from natural scenes. Each basis function in the dictionary is a $8 \times 8 = 64$D vector. A sample learned dictionary of size 200 is shown in Fig. 3. We experimented with different dictionary sizes (10, 50, 100, 200, 400, and 1000) and realized that fixation prediction results did not change much. The sparse codes $\boldsymbol{\alpha}_i$ are computed with the above basis using the LARS algorithm [5] implemented in the SPAMS toolbox[1].

## 2.2. Measuring visual saliency

Our model is based on two saliency operations. The first one, local contrast, considers the rarity of image regions with respect to (small) local neighborhoods (guided by the well-established computational architecture of Koch and Ullman [19] and Itti *et al.* [4]). The second operation, global contrast, evaluates saliency of an image patch using its contrast with respect to the patch statistics over the entire image. Finally, local and global contrast maps are consolidated. We repeat the process for each channel of both RGB and Lab color systems and fuse saliency maps of each sub channel of a color space to generate a saliency map for each color system. At each stage, maps are normalized before integration (See Fig. 2).

**Local saliency.** Local saliency ($S_l$) in our model is the average weighted dissimilarity between a center patch $i$ (blue rectangle in Fig. 4) and its $L$ patches in a rectangular neighborhood (red rectangle in Fig. 4):

$$S_l^c(\mathbf{p}_i) = \frac{1}{L} \sum_{j=1}^{L} W_{ij}^{-1} D_{ij}^c \quad (3)$$

where $W_{ij}$ is the Euclidean distance between the center patch $i$ and the surround patch $j$. Thus, those patches further away from the center patch will have less influence on

[1] http://www.di.ens.fr/willow/SPAMS/index.html

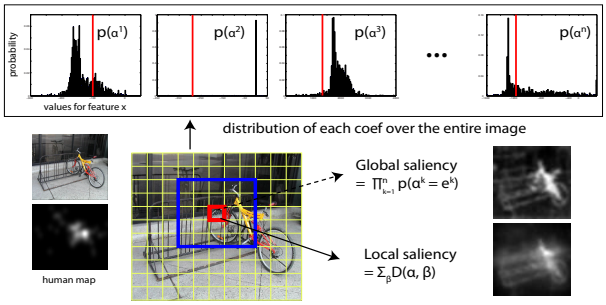



Figure 4. **Illustration of global and local saliency for an image patch.** Global saliency measures the rarity of patch in the entire scene while local rarity measures the difference between a patch and its surrounding context.

the saliency of the center patch. $D_{ij}$ denotes the Euclidean distance between patch $i$ and patch $j$ in the feature space between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$, vectors of coefficients for patches $i$ and $j$, respectively derived from sparse coding (Sec. 2.1). While here we use the Euclidean distance ($l_2$ distance), the KL distance [23][38], $l_1$ distance [17], or correlation coefficient have also been used in the past to calculate patch similarity. Superscript $c$ denotes color sub channels ($L, a, or\ b$ in $Lab$ or $R, G, or\ B$ in $RGB$).

**Global saliency.** It often happens that a local patch is similar to its neighbors but the whole region (i.e., local + surrounding) is still in global rarity in the entire scene. Using only the local saliency may suppress areas within a homogeneous region resulting in blank holes, which sometimes impedes object-based attention (e.g., a uniformly textured object would only be salient at its borders). To remedy such shortcoming, we build our global saliency operator guided by the information-theoretic saliency measure of Bruce and Tsotsos [14]. Instead of each pixel, here we calculate the probability of each patch $P(\mathbf{p}_i)$ over the entire scene and use its inverse as the global saliency:

$$\begin{aligned} S_g^c(\mathbf{p}_i) &= P(\mathbf{p}_i)^{-1} = \left(\Pi_{j=1}^{n} P(\alpha_{ij})\right)^{-1} \\ log\left(S_g^c(\mathbf{p}_i)\right) &= -log(P(\mathbf{p}_i)) = -\sum_{j=1}^{n} log(P(\alpha_{ij})) \\ S_g^c(\mathbf{p}_i) &\propto -\sum_{j=1}^{n} log(P(\alpha_{ij})) \end{aligned} \quad (4)$$

To calculate $P(\mathbf{p}_i)$, we assume that coefficients $\boldsymbol{\alpha}$ are conditionally independent from each other. This is to some extent guaranteed by the sparse coding algorithm [5]. For each coefficient of the patch representation vector (i.e., $\alpha_{ij}$), first a binned histogram (100 bins here) is calculated from all of the patches in the scene and is then converted to a pdf ($P(\alpha_{ij})$) by dividing to its sum. If a patch is rare in one of the features, the above product will get a small value leading to high global saliency for that patch overall. Fig. 4 illustrates the process of calculating global saliency.

**Combined saliency.** Local and global saliency maps are

then normalized and combined:

$$S_{lg}^{c}(\mathbf{p}_i) = \mathcal{N}(S_{l}^{c}(\mathbf{p}_i)) \circ \mathcal{N}(S_{g}^{c}(\mathbf{p}_i)) \quad (5)$$

where $\circ$ is an integration scheme (i.e., $\{+, *, \max, or \min\}$). Through the experiments, we found that $max$ in this stage leads to slightly higher accuracy than others. Then, saliency values of a patch in all channels are normalized and summed again to generate the saliency of a patch in each color system. For Lab color system, we have:

$$S_{lg}^{Lab}(\mathbf{p}_i) = \sum_{c \in L,a,b} \mathcal{N}(S_{lg}^{c}(\mathbf{p}_i)) \quad (6)$$

The same operation applies to the RGB color space. Final saliency for a patch is then summation of normalized saliency maps in both color systems:

$$S_{lg}(\mathbf{p}_i) = \mathcal{N}(S_{lg}^{Lab}(\mathbf{p}_i)) + \mathcal{N}(S_{lg}^{RGB}(\mathbf{p}_i)) \quad (7)$$

**Normalization ($\mathcal{N}$).** Similar to [4] first, the average of all local maxima (defined as greater than 4 neighboring points) with intensity above a threshold is calculated ($M_l$). Then a map is multiplied by $p = (M_g - M_l)^2$ where $M_g$ is the global maximum in the map (known as maxnorm).

**Extension to the scale space.** Since objects appear at different sizes and depths, it is necessary to perform saliency detection at several spatial scales. To make our approach multi-scale, we calculate the saliency of downsampled images (divisions by 2) from the original image and then take the average after normalization:

$$S(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \mathcal{N}(S_{lg}^{i}(\mathbf{x})) \quad (8)$$

where $M$ is the number of scales and $S_{lg}^{i}(\mathbf{x})$ is the saliency of pixel $\mathbf{x}$ derived from the saliency map created by Eq. 7. Finally, we smooth the resultant map by convolving it with a small Gaussian kernel for better visualization.

**Handling center-bias.** A tedious and challenging factor in saliency modeling is handling center-bias in eye tracking data, which is the tendency of human subjects to preferentially look near the image center [13]. This generates a high central peak in the overall 2D histogram of fixations, resulting in high scores for a trivial saliency model whose map is just a Gaussian blob at the image center. To account for center-bias, some models intrinsically (e.g., GBVS [7], E-Saliency [37]) or extrinsically (e.g., Judd *et al.* [3] and Yang *et al.* [16]) add a center prior to their algorithms[2]. Here, instead of adding center bias to our model, we use a scoring metric that discounts center-bias in a non-parametric manner (See next section) when evaluating our and other saliency models against eye-tracking data.

[2]Some models add center-bias by either fitting a 2D Gaussian to fixation data or simply by just using the average fixation map (i.e., 2D histogram).

## 3. Experimental Setup

To validate our proposed method, we carried out several experiments on four benchmark datasets using the "shuffled AUC" score described below[3]. The main reason behind employing several datasets is that current datasets have different image and feature statistics, stimulus variety, biases (e.g., center-bias), and eye tracking parameters. Hence, it is necessary to employ several datasets as models leverage different features that their distribution varies across datasets.

**Evaluation metric.** The most widely used score for saliency model evaluation is the **AUC** [14]. In AUC, human fixations for an image are considered as the positive set and some points from the image are randomly chosen (uniformly) as the negative set. The saliency map is then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map and plotting true positive rate vs. false positive rate, an ROC curve is achieved and its underneath area is calculated. A problem with AUC is that it generates a large value for a central Gaussian model and is thus affected by center-bias [13]. To tackle center bias, Zhang *et al.* [12] introduced **shuffled AUC** score, with the only difference that instead of selecting negative points randomly from a uniform distribution, all human fixations (except the positive set) are used as the negative set. Shuffled AUC score generates a value of 0.5 for both a central Gaussian and a completely uniform map. Please note that in addition to shuffled AUC, there are also some other scores that have been often used in the past, for example Normalized Scanpath Saliency (NSS) [35], KL distance [10], and Correlation Coefficient [46]. But here we avoid using them as they are all affected by center-bias. Instead, we adopt the shuffled AUC score which is becoming a standard for saliency model evaluation [40][12].

Utilized fixation datasets are briefly described below.

**TORONTO[4] [14].** This is the most widely used dataset for model comparison. It contains 120 color images with resolution of $511 \times 681$ pixels from indoor and outdoor environments. Images are presented at random to 20 subjects for 3 seconds with 2 seconds of gray mask in between.

**MIT[5] [3].** This is the largest dataset containing 1003 images (resolution from $405 \times 1024$ to $1024 \times 1024$ pixels) collected from Flicker and LabelMe datasets. There are 779 landscape and 228 portrait images. Fifteen subjects freely viewed images for 3 sec. with 1 sec. delay in between.

**KOOTSTRA[6] [20].** This dataset contains 101 images from 5 different categories: 12 animals, 12 automan, 16 buildings, 20 flowers, and 41 natural scenes. Images are ob-

[3]Our software for score calculation and saliency maps over 4 datasets are available at: https://sites.google.com/site/saliencyevaluation/.

[4]Available at: http://www-sop.inria.fr/members/Neil.Bruce

[5]This dataset is available at: http://people.csail.mit.edu/tjudd/

[6]This dataset is available at: http://www.csc.kth.se/~kootstra/

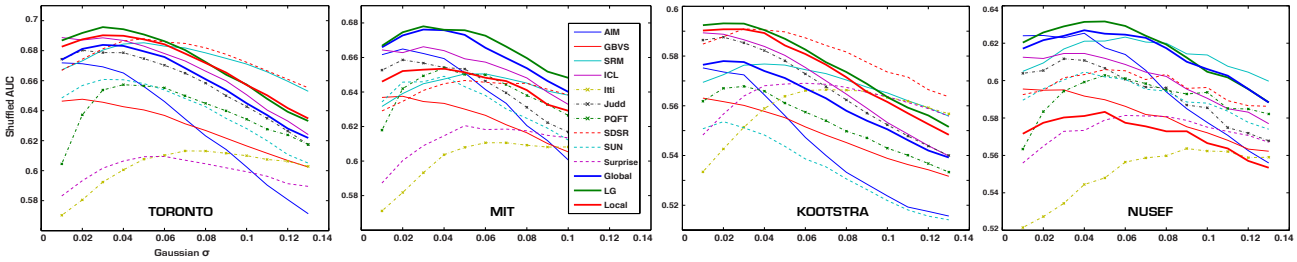



Figure 5. **Model comparison.** Fixation prediction accuracy of our saliency operations (Local, Global, LG (Local + Global)) along with 10 state-of-the-art models over 4 benchmark datasets. X-axis indicates the $\sigma$ of the Gaussian kernel (in image width) by which maps are smoothed. NUSEF dataset contains some images with copyright which are not easily accessible and we don't use. Only 412 images are used here.

| Dataset | AIM [14] | GBVS [7] | SRM [8] | ICL [9] | Itti [4] | Judd [3] | PQFT [24] | SDSR [11] | SUN [12] | Surprise [10] | Local $S_l$ | Global $S_g$ | LG $S_{lg}$ | Gauss | IO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TORONTO [14] | 0.67 | 0.647 | 0.685 | **0.691** | 0.61 | 0.68 | 0.657 | 0.687 | 0.66 | 0.605 | **0.691** | 0.69 | **0.696** | 0.50 | **0.73** |
| Optimal $\sigma$ | 0.01 | 0.02 | 0.05 | 0.01 | 0.07 | 0.03 | 0.04 | 0.05 | 0.03 | 0.06 | 0.04 | 0.03 | 0.03 | - | - |
| MIT [3] | 0.664 | 0.637 | 0.65 | **0.666** | 0.61 | 0.658 | 0.65 | 0.646 | 0.649 | 0.62 | 0.653 | **0.676** | **0.678** | 0.50 | **0.75** |
| Optimal $\sigma$ | 0.02 | 0.02 | 0.05 | 0.03 | 0.06 | 0.02 | 0.04 | 0.05 | 0.04 | 0.05 | 0.04 | 0.04 | 0.03 | - | - |
| KOOTSTRA [20] | 0.575 | 0.563 | 0.576 | 0.589 | 0.57 | 0.587 | 0.57 | **0.59** | 0.55 | 0.566 | **0.591** | 0.578 | **0.593** | 0.50 | **0.62** |
| Optimal $\sigma$ | 0.01 | 0.01 | 0.04 | 0.01 | 0.07 | 0.02 | 0.03 | 0.03 | 0.02 | 0.07 | 0.03 | 0.02 | 0.03 | - | - |
| NUSEF [6] | **0.623** | 0.595 | 0.62 | 0.614 | 0.56 | 0.61 | 0.60 | 0.60 | 0.60 | 0.58 | 0.583 | **0.627** | **0.632** | 0.49 | **0.66** |
| Optimal $\sigma$ | 0.04 | 0.01 | 0.06 | 0.03 | 0.09 | 0.03 | 0.05 | 0.04 | 0.04 | 0.06 | 0.05 | 0.04 | 0.05 | - | - |

Table 1. **Maximum performance of models shown in Fig. 5**. Numbers in second rows are the sigma values where models take their maximum performance. Parameter settings: Surround window size = 1; number of scales = 1 ($256 \times 256$). Accuracies of three best models over each dataset are shown in bold face font. LG is the Local+Global model and IO stands for the human inter-observer model.

served by 31 subjects in the age range of 17 to 32 for 5 seconds. Image resolution is $768 \times 1024$ pixels. This dataset is specially challenging because there are not explicit objects or salient regions within many of the images.

**NUSEF**[7] **[6].** This dataset includes 758 images containing emotionally affective scenes/objects such as expressive faces, nudes, unpleasant concepts, and interactive actions. In total, 75 subjects free-viewed part of the image set for 5 seconds each (on average 25 subjects per image).

## 4. Performance Evaluation

Here, along with the evaluation of our model, we also compare 10 state-of-the-art bottom-up saliency models. Softwares for these models are publicly available[8]. Additionally, we implemented two simple models, to serve as baseline: Gaussian Blob (Gauss) and Human inter-observer (IO). Gaussian blob is simply a 2D Gaussian shape drawn at the center of the image; it is expected to predict human gaze well if such gaze is strongly clustered around the center [13]. The human model outputs, for a given stimulus, a map built by integrating fixations from other subjects than the one under test while they watched that stimulus. The human map is usually smoothed by convolving with a small Gaussian kernel. This model provides an upper-bound on prediction accuracy of saliency models to the degree that, different humans may be the best predictors of each other. Model maps were resized to the size of the original image, onto which eye data have been recorded.

An important parameter in model comparison is smoothness (blurring) of saliency maps [40]. Here, we smoothed the saliency map of each model by convolving it with a variable size Gaussian kernel. Fig. 5 presents the shuffled AUC score of models over the range of standard deviations $\sigma$ of the Gaussian kernel in image width (from 0.01 to 0.13 in steps of 0.01). The maximum score value over this range for each model is shown in Table 1. Smoothing the saliency map dramatically affects the accuracy of some models (e.g., Itti, Surprise, PQFT, and SUN). Our combined saliency model (local + global and RGB + Lab) outperforms other models over 4 datasets with a larger margin over the MIT and NUSEF datasets. Our local and global saliency operators have less accuracy than the combined model but are still above several models. Results show that global saliency works better than local saliency operator over large datasets (MIT and NUSEF) while they are close to each other over TORONTO dataset. Models were more successful over the TORONTO and MIT datasets and less over KOOTSTRA and NUSEF, possibly because of the higher complexity of stimuli in these datasets. The NUSEF dataset contains many affective and emotional stimuli while KOOTSTRA dataset contains images without well-defined interesting and salient objects (e.g., nature scenes, trees, and flowers).

[7] Available at: http://mmas.comp.nus.edu.sg/NUSEF.html

[8] **AIM**: http://www-sop.inria.fr/members/Neil.Bruce/
**GBVS**: http://www.klab.caltech.edu/~harel/
**SRM & ICL**: http://www.klab.caltech.edu/~xhou/
**Itti & Surprise**: http://ilab.usc.edu/toolkit/
**Judd**: http://people.csail.mit.edu/tjudd/
**PQFT**: http://visual-attention-processing.googlecode.com/
**SDSR**: http://alumni.soe.ucsc.edu/~rokaf/
**SUN**: http://cseweb.ucsd.edu/~l6zhang/

| Dataset | RGB | | | Lab | | | RGB + Lab | | |
|---|---|---|---|---|---|---|---|---|---|
| | $S_l$ | $S_g$ | $S_{lg}$ | $S_l$ | $S_g$ | $S_{lg}$ | $S_l$ | $S_g$ | $S_{lg}$ |
| TORONTO | 0.646 | 0.647 | 0.653 | 0.670 | 0.660 | 0.660 | 0.678 | 0.668 | **0.683** |
| MIT | 0.627 | 0.639 | 0.640 | 0.646 | 0.644 | 0.651 | 0.658 | 0.663 | **0.667** |
| KOOTSTRA | 0.574 | 0.572 | 0.578 | 0.572 | 0.555 | 0.570 | 0.589 | 0.573 | **0.591** |
| NUSEF | 0.599 | 0.610 | 0.610 | 0.556 | 0.596 | 0.592 | 0.569 | 0.614 | **0.616** |

Table 2. **RGB vs. Lab for saliency detection.** $S_l$: Local; $S_g$: Global; $S_{lg}$: Local + Global. Parameter settings: scales ($M$) =1 ($256 \times 256$); Window size = 1. Results are over original saliency maps without smoothing.

Among compared models, ICL [12], AIM [14], SDSR [11], and Judd *et al.* [3] performed higher than the rest. Itti *et al.* [4] and Surprise [10] models are ranked at the bottom. As we expected, a trivial Gaussian blob located at the image center scores around 0.5 over all datasets and human model scores the best providing a gold standard for visual saliency models. Humans are less correlated over KOOTSTRA and NUSEF datasets.

**Lab vs. RGB for saliency detection.** To assess the power of Lab and RGB color spaces for saliency detection, we performed an experiment using each of the two color systems. Results over all four datasets are shown in Table 2. According to our results, it is not possible to tell which color system is the best. The Lab color space leads to higher accuracies over TORONTO and MIT datasets while RGB works better over KOOTSTRA and NUSEF datasets. Integrating both color systems leads to higher accuracy than each one taken separately, consistently over all four datasets. This indicated the importance of saliency integration over both color systems. Also note that over each component (local or global), combination of RGB and Lab leads to higher performance than each of the color systems.

**Influence of surrounding window size and number of scales.** Here we analyze how the size of the surrounding window (and hence number of neighbors) and number of spatial scales affect performance of our models. As the left diagram of Fig. 6 shows, increasing the number of neighbors reduces the accuracy of the local saliency operator. Correspondingly, this reduces the accuracy of the combined model. Note that the global operator is not affected by this parameter. Shown in the right panel of Fig. 6, increasing the number of scales enhances the results to a certain point (here using 3 scales [256, 128, and 64]) and then drops (when using 4 scales [512 256 128 64]).

**Runtime aspects.** It takes approximately 5 seconds for our model to process a $256 \times 256$ image in both RGB and Lab color spaces using a personal computer running Linux Ubuntu with 5.8 GB RAM and 12 Core Intel i7 3.2 GHz CPU. Our model is faster than AIM (16 sec), Judd (without object detectors)(4.7 sec), close to SDSR (2.4) and GBVS (2), and slower than PQFT (1 sec), SUN (1), Itti (0.28) (using [52]), and ICL (0.1) models. Our global saliency operator is appx. 3 times faster than our local saliency operator.

For qualitative assessment, we show in Fig. 7 saliency maps of our combined saliency model and compared models for sample images from TORONTO and MIT datasets.

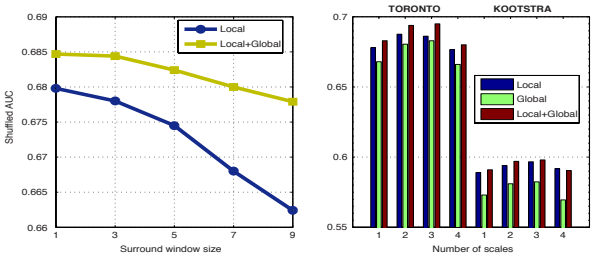



Figure 6. **Parameter analysis**. Left: Effect of the surround window size on accuracy over TORONTO dataset using $256 \times 256$ images ($M = 1$). Right: Influence of scale on results over TORONTO and KOOTSTRA datasets (window size =1). First three bars are 256,128,64 and fourth one represents four scales 512,256,128,64.

## 5. Conclusions and Future Works

We enhance the state-of-the-art in saliency modeling by proposing an accurate and easy-to-implement model that utilizes image representations in both RGB and Lab color spaces. Furthermore, we introduce one local and one global saliency operator each representing a class of previous models to some extent. We conclude that integration of local and global saliency operators works better than just using either one, which encourages more research in this direction. Similarly, combining both color systems strongly benefits saliency detection and eye fixation prediction.

There are two areas that we would like to improve upon. The first one is incorporating top-down factors for fixation prediction. The large gap between models and the human inter-observer model (see Table 1) is mainly due to role of top-down concepts (e.g., faces, text [45], people, and cars [3], affective and emotional stimuli or actions within scenes [6]) when freely viewing scenes. While some of these factors have been utilized for saliency detection in the past [3], adding more top-down features (e.g., by reliable detection of text on natural scenes) can scale up accuracy of current models. The second area is extending our model for saliency detection in spatio-temporal domain (videos).

*Supported by the National Science Foundation (grant number BCS-0827764), and the Army Research Office (W911NF-08-1-0360 and W911NF-11-1-0046), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.*

## References

[1] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.*, 5:1-7, 2004. 1

[2] W., Kienzle, A. F., Wichmann, B., Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. *NIPS*, 2007. 2

[3] T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look, *ICCV*, 2009. 2, 3, 5, 6, 7

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 1998. 1, 2, 3, 4, 5, 6, 7

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. of Machine Learning*, 2010. 4

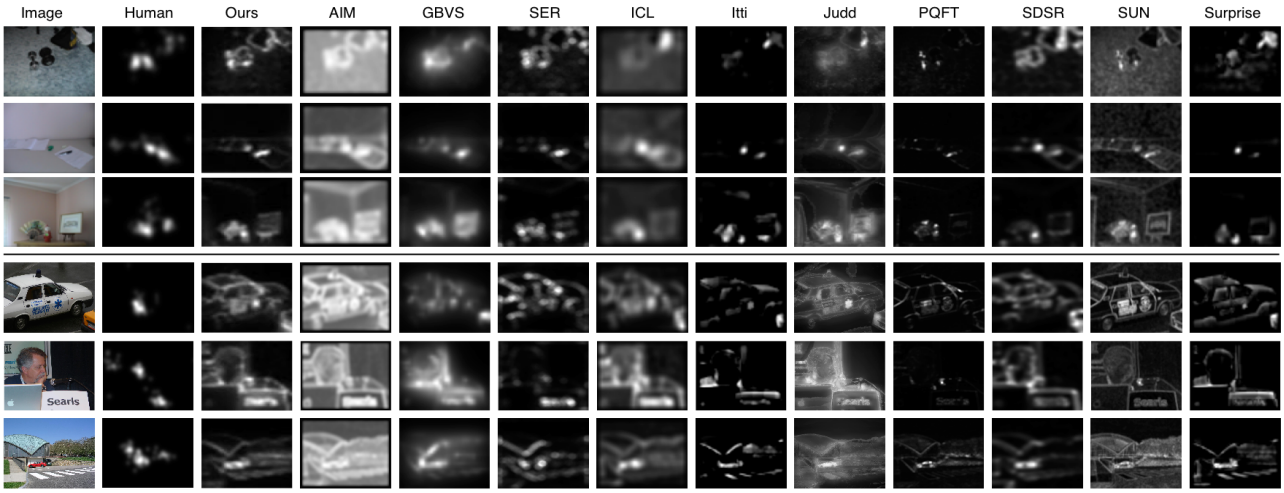



Figure 7. **Visual comparison of our combined saliency model and 10 state-of-the-art models over samples from TORONTO (top) and MIT datasets.**


[6] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. *ECCV*, 2010. 6, 7

[7] J. Harel, C. Koch, P. Perona. Graph-based visual saliency. *NIPS*, 2006. 2, 5, 6

[8] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007. 2, 6

[9] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 2008. 2, 3, 6

[10] L. Itti and P. Baldi. Bayesian surprise attracts human visual attention. *NIPS*, 2005. 2, 5, 6, 7

[11] H.J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9, 2009. 2, 6, 7

[12] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, SUN: A Bayesian framework for saliency using natural statistics. *J. of Vision*, 8(32):1-20, 2008. 2, 3, 5, 6, 7

[13] B.W. Tatler. *J. Vision*, 14(7):1-17, 2007. 5, 6

[14] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005. 1, 2, 3, 4, 5, 6, 7

[15] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psych.*, 12:97-136, 1980. 1, 3

[16] Y. Yang, M. Song, N. Li, J. Bu, and C. Chen. What is the chance of happening:A new way to predict where people look. ECCV,2010. 5

[17] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. CVPR 2011. 2, 4

[18] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *PAMI*, 2006. 2

[19] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 1985. 1, 3, 4

[20] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *BMVC*, 2008. 2, 5, 6

[21] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. . *ICCV*, 2009. 2

[22] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011. 2

[23] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. *NIPS*, 2007. 2, 4

[24] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and Its applications in image and video compression. *IEEE Trans. on Image Processing*, 2010. 2, 6

[25] M. Mancas. Computational attention: Modelisation and application to audio and image processing. PhD. thesis, 2007. 2

[26] A. Torralba, A. Oliva, M. Castelhano and J.M. Henderson. Contextual guidance of attention in natural scenes: The role of Global features on object search. *Psychological Review*, 2006. 2

[27] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996. 3

[28] E. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24, 2001. 3

[29] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. *CVPR*, 2006. 2

[30] W. Einhäuser, M. Spain, and P. Perona. Objects predict xations better than early saliency. *Journal of Vision*, 2008. 3

[31] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using national image. *CVPR*, 2010. 3

[32] F. Bach, J. Mairal, J. Ponce, and G. Sparo. Sparse coding and dictionary learning for image analysis. *CVPR*, 2010. 3

[33] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma. A review of fast l1-minimization algorithms for robust face recognition. *http://arxiv.org*, 2010. 3

[34] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3336-3745, 2006. 3

[35] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Res.*, 45, 2005. 5

[36] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Res.*, 2002. 2

[37] T. Avraham, M. Lindenbaum. Esaliency (Extended Saliency): Meaningful attention using stochastic image modeling. *PAMI*, 2010. 5

[38] D.A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. *ICCV*, 2011. 4

[39] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilá. Segmenting salient object from images and videos. *ECCV*, 2010. 2

[40] X. Hou, J. Harel, and Christof Koch. Image Signature: Highlighting sparse salient regions. *IEEE PAMI*, In press. 5, 6

[41] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. *SPIE*, 2003. 2

[42] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. *ACIVS*, 5807, 2009. 2

[43] L. Elazary and L. Itti. Interesting objects are visually salient. *J. Vision*, 2008. 3

[44] M.M Cheng, G.X Zhang, N.J. Mitra, and X. Huang, and S.M. Hu. Global Contrast based Salient Region Detection. *CVPR*, 2011. 2

[45] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting gaze using low-level saliency combined with face detection. *NIPS*, 2007. 3, 7

[46] N. Ouerhani, R. von Wartburg, H. Hugli, and R.M. Muri. Empirical validation of saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 2003. 5

[47] A. Toet. Computational versus psychophysical image saliency: A comparative evaluation study. *IEEE trans. PAMI*, 2011. 1

[48] N. Parikh, L. Itti, and J. Weiland. Saliency-based image processing for retinal prostheses. *J. Neural Eng.* 7, 2010. 1

[49] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans. Image Process.*, 2004. 1

[50] J. Wang, J. Sun, L. Quan, X. Tang, and H.Y Shum. Picture collage. *CVPR*, 1:347-354, 2006. 1

[51] R. Rosenholtz, A. Dorai, and R. Freeman. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)*, 2011. 1

[52] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 2006. 1, 7

[53] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Trans. on Graphics*, 2002. 1

# Probabilistic Learning of Task-Specific Visual Attention

Ali Borji Dicky N. Sihite Laurent Itti
Department of Computer Science, University of Southern California, Los Angeles
http://ilab.usc.edu

## Abstract

*Despite a considerable amount of previous work on bottom-up saliency modeling for predicting human fixations over static and dynamic stimuli, few studies have thus far attempted to model top-down and task-driven influences of visual attention. Here, taking advantage of the sequential nature of real-world tasks, we propose a unified Bayesian approach for modeling task-driven visual attention. Several sources of information, including global context of a scene, previous attended locations, and previous motor actions, are integrated over time to predict the next attended location. Recording eye movements while subjects engage in 5 contemporary 2D and 3D video games, as modest counterparts of everyday tasks, we show that our approach is able to predict human attention and gaze better than the state-of-the-art, with a large margin (about 15% increase in prediction accuracy). The advantage of our approach is that it is automatic and applicable to arbitrary visual tasks.*

## 1. Introduction

Visual attention is an important facet of our vision in everyday life. It makes processing complex visual scenes tractable through sequential selection of localized image regions. It is commonly believed that visual attention is guided by two components: 1) a bottom-up (BU), task-independent, and image-based component that instinctively draws the eyes to places in the scene that contain discontinuities in image features, such as motion, color, and texture, and 2) a top-down (TD) component that guides attention and gaze in a task-dependent and goal-directed manner, orchestrating the sequential acquisition of information from the visual environment. In everyday life, these two components are combined in the control of gaze.

In computer vision, research on visual attention has been primarily focused on the BU component. Early studies were directly influenced by cognitive studies of visual search and Feature Integration Theory (FIT) [12]. This led Koch and Ullman [13] to define the saliency map: A topographic map with retinotopic organization where locations that stand out in an image (e.g., because of distinctive features such as color, texture, and motion) are highlighted. The first complete implementation and verification of this architecture was done by Itti *et al.* [7]. Several other approaches for detecting image-based outliers have also been proposed, based on information theory [14], discriminant hypothesis [15], spectral models [16], sparse and efficient coding [17], and Bayesian and graphical models [18][19][21].

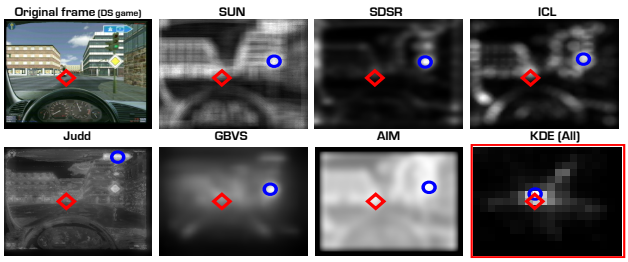



Figure 1. Bottom-up saliency does not account for task-driven eye movements. Predictions of 6 state-of-the-art BU saliency models in a driving scene and our model (red box). Red diamond and blue circle show human fixation and maximum of a model, respectively. See Table 2 for results.

Today, saliency detection and eye movement prediction over static images and videos is a reasonably well-researched area and there are many models with good accuracy, although of course improving tolerance to noise and invariance of algorithms is always possible. However, success of BU models is limited to a small range of everyday tasks, such as free-viewing [7][14][18] and their adaptation to visual search [22][11]. BU models usually can not predict exact fixation points and leave more than half of them unaccounted [25] (Fig. 1). One problem with models based on saliency maps is that they are correlated with fixation behaviors but don't tell much about the cause for such behaviors [26][27]. Diverging from the current trend, we focus on modeling top-down attention which can boost performance of several approaches in computer vision. For example in areas such as object detection and recognition [15][35][36], especially in spatio-temporal domain for video understanding and action/event recognition (e.g., [37][38]).

We aim to build an attentive vision system that can tell where it should look as it moves through the world and interacts with the environment. This problem is very important but very difficult and largely unsolved. Our approach is to utilize global visual context [28][18], a low-dimensional representation of the whole image (the "gist" of the scene).

Such a representation can be easily computed, and it relaxes the need to identify specific regions or segment and recognize all objects in a scene. We focus on interactive environments (contemporary 2D and 3D video games), where visual stimuli are dynamically generated and affected by deliberative motor actions. We develop top-down models trained over the same or similar games from data of subjects during game playing, and we use those models to predict saccades of a new test subject. Several sources of multimodal information, including global context, previous saccades, and previous motor actions, are combined in a unified Bayesian framework over time. Compared with brute-force algorithms such as the average of all saccade positions, a central Gaussian blob, and 6 popular BU models, we show that our models significantly outperform the state-of-the-art in terms of accuracy at predicting where a new subject looks during active gameplay. This indicates the effectiveness of our approach for modeling complex task-driven attention.

**Previous Work.** The majority of studies on TD attention are at the analysis/descriptive level and there are few computational models available, we believe due to conceptual complexity. Yarbus [1] discovered a compelling finding that 'seeing' is inextricably linked to the observer's cognitive goals. Task dependency of gaze has been extensively studied for several real-world tasks, such as "sandwich making" [4], "tea making" [2], and "driving" [5]. These studies have revealed that most fixations are directed to task-relevant locations, and there is a tight temporal relationship between fixations and task-related behaviors, such that it is sometimes possible to infer the algorithm of a task from the pattern of a subject's eye movements (e.g., in "block copying" [4]). In [3], Hayhoe and Ballard elaborate on the role of internal reward in guiding eye and body movements, supported by neurophysiological studies. Inspired by the idea of visual routines [29] and using reinforcement learning (RL) approaches, Sprague and Ballard [9] proposed an RL-based top-down attention model for explaining eye movements of an agent operating in virtual environments. This approach is interesting but suffers from three limitations that make it hard to apply directly for computer vision purposes. First, it is limited to laboratory-scale tasks such as side-walk navigation [9], second, visual processing is very simple, and, third, it needs explicit definitions of reward functions, subtasks, and arbitration mechanisms.

Our approach has in part similarities with the contextual model of Torralba *et al.* [18] as we also use the concept of gist. We start with a basic Bayesian formulation and add new features to account for task-driven attention in spatio-temporal domain while former has been thus far utilized for bottom-up saliency and visual search over static stimuli. This model and its decedents (e.g., [19, 11]) originally formulate object search as estimating the probability $P(O=1, X|L, G)$ where $X=(x, y)$ defines the location of the target in the image, $O$ is a binary variable ($O=1$ denotes target presence and $O=0$ denotes target absence in the image), and $L$ and $G$ denote local and global features, respectively. According to Bayes' theorem, they expand the above probability as:

$$P(O=1, X|L, G) = \frac{1}{P(L|G)} P(L|O=1, X, G) P(X|O=1, G) P(O=1|G) \quad (1)$$

The first term on the right, $\frac{1}{P(L|G)}$, is independent of the target, measures BU saliency, and is solely dependent of local image features. The second term represents top-down knowledge of target appearance. Image regions with features likely to belong to the target object are enhanced. The third term provides context-based priors on the location of the target, and the fourth term provides the prior probability of presence of the target in the scene. If this probability is very small, then object search need not be initiated. Zhang *et al.* [19] used Independent Component Analysis (ICA) and Difference of Gaussians (DOG) features learned from a large repository of natural scenes to estimate the first term. From another perspective these models unify the information theoretic models (e.g., [14]), in the sense both are based on self-similarity of scene regions. These models assign higher saliency values to regions with rare features. Information of visual feature $F$ is $I(F) = -log\ P(F)$ which is inversely proportional to the likelihood of observing $F$. By fitting a distribution $P(F)$ to features, rare features can be immediately found by computing $P(F)^{-1}$ in an image. The idea of global context has also been extensively employed in several areas of computer vision (e.g., [32][10]).

Several other approaches have been proposed to model top-down attention, specifically for visual search. Navalpakkam and Itti [22] proposed a Bayesian approach to derive the optimal gains that should be applied to low-level visual features contributing to a saliency model [7], to make an object of interest more salient. The objective was to maximize the signal to noise ratio of the expected target object versus background clutter, and training was performed over a set of natural scenes containing ground-truthed objects. An intuitive solution for the same problem (optimal gains of feature channels) was suggested earlier by Frintrop [23] which is the end result of the SNR maximization process in [22]. Navalpakkam and Itti [8] proposed conceptual guidelines for modeling the role of task on visual attention, but their method requires the algorithm of the task to be known, and is not fully implemented.

Perhaps the most similar work to ours (i.e., real-world and unconstrained tasks) is the work by Peters and Itti [6], where they used gist as a predictor of fixation, learning from examples where people looked in scenes of different gists and while engaged in a particular task. The same scene gist, however, might not always warrant the same eye movement, based on the history and sequence of previous fixations and

actions to date. For example, in one of the games studied here, even when looking at the exact same scene, eye movements are often guided by past events, such as different customers placing different orders for items which the player is asked to provide. To tackle the problem that gist of the scene is not enough, we follow a sequential processing framework where several factors predictive of eye movements are integrated over time and can resolve the confusion (aliasing) at one snapshot of time.

## 2. Proposed Model

Our goal is to predict where a human subject attends under the task influence $T$. This is similar to explaining saccades (jumps in eye movements) in free-viewing, addressed by bottom-up models, with the difference that here a policy governs saccades. Since it is difficult to learn general strategies for performing every task, here we focus on learning models for each task separately. Following a leave-n-out approach over subjects, first, in the training phase, we compile a training set of feature vectors and eye positions corresponding to individual frames from several video game clips which were recorded while observers were playing video games. Then, training data is used to learn probability distributions over image locations for given feature vectors, and pdfs are later leveraged in the test phase for inferring the next attended location of a new test subject.

We need a number of variables that cause or correlate with saccade positions and hence can provide information regarding the next saccade location. These variables tell us indirectly about the state of the agent at each time point of the task. In addition to scene gist, here, we introduce two new features explained below: motor actions and previous saccade position and then combine them in a probabilistic manner over time to infer a probability distribution over scene locations that may attract next saccade:

**Global context (Gist, $G$).** Following a brief presentation of a photograph, humans are able to summarize the quintessential characteristics of an image, a process previously expected to require much analysis. A number of models exist for calculating Gist (e.g., [18][33]). We adopt the gist model of [10][1] as it is based on the bottom-up saliency model [7] that we use here as a baseline approach. We consider 4 scales for each orientation pyramid, 6 scales for each color pyramid, and 6 scales for intensity. For each of the maps, average in each of the patches of grid sizes $n \times n$ (here $n \in \{1, 2, 4\}$) are calculated (thus 21 values). Overall the final gist vector will be augmentation of $(4 \times 4 + 6 \times 2 + 6 \times 1) \times 21 = 714$ values. We then employ PCA to reduce the dimensionality. We also, investigate the ability of histogram of oriented gradient (HOG) [30] features to represent the global context of a scene[2].

[1] http://ilab.usc.edu/siagian/Research/Gist/Gist.html

[2] http://pascal.inrialpes.fr/soft/olt/

**Previous saccade location ($X$).** A lot of everyday tasks need a number of perceptions and actions to be performed in a sequence (e.g., sandwich making [4]). Therefore, knowing what object has been attended previously gives an evidence for the next attended object. We implement this idea over spatial locations. For instance, $P(X^{t+1} = b|X^t = a)$ indicates the probability of looking at location $b$ in the next time step given that location $a$ is currently fixated (e.g., looking at left first and then right, when turning right).

**Motor actions ($A$).** Actions and fixations are tightly linked thus, by knowing a performed action, one can tell where to look next. We recorded motor actions while humans were involved in game playing. We assumed that these actions correspond to some high-level events in the game (e.g., mouse click for shooting). We logged actions for driving games (e.g., wheel position, pedals (brake and gas), left and right signals, mirrors, left and right side views, and gear change), from which we only generated a 2D feature vector from wheel and pedal positions between 0 and 255 (Fig. 2). For other games, 2D mouse position and joystick buttons were used (further explained in Sec. 3.1).

### 2.1. Problem Formulation

In this section, we describe details of our Bayesian approach of information integration over time to predict saccade in the next time step. Our method is based on Hidden Markov Models (HMM), which are successful probabilistic tools for sequence processing. We are particularly interested in the probability of attending to spatial location $X$ given all available information $I$, or $P(X|I)$. One way to estimate this probability is to follow a discriminative approach by augmenting all information into a large vector $I$, and using a classifier to map it to $X$ from a set of labeled training data. An alternative is to follow a Bayesian formulation: $P(X|I) = P(I|X)P(X)/P(I) = \mu P(I|X)P(X)$. Parameter $\mu$ is selected in a way that resultant probabilities sum to 1 (i.e., $\sum_j P(X_j|I) = 1$). $P(X)$ is simply the prior distribution of all saccade locations in the training data (sum of all saccades or average fixation map). A benefit of the generative approach over the discriminant classifier-based approach is that, it provides a unified method for information integration of sequential data, and makes it suitable for our purpose. which enhances results.

Formally, the goal of the saccade prediction is to compute a probability distribution over the possible locations given all features up to time $t$. Let $X_t \in \{1 \dots n\}$ denote the saccade location with $n$ as the number of locations in the image at time $t$. To generate sufficient data, we resize the original eye fixation map with one at the attended location and zeros elsewhere into a smaller scale map (a $w \times h$ grid). Therefore, $X_t$ is the location of 1 in such map. In the following we start with the simplest case of $P(X|I)$ when only global context information is available (i.e., $I$ is equal to Gist) and add more information in subsequent steps.

**Case 1: Gist only.** In this case, only global context information from all past and the current time is used. According to the Bayes theorem we have:

$$\begin{aligned} P(X_t|G_{1:t}) &= P(X_t|G_t, G_{1:t-1}) \\ &= \frac{P(G_t|X_t)P(X_t|G_{1:t-1})}{P(G_t|G_{1:t-1})} \\ &= \mu P(G_t|X_t)P(X_t|G_{1:t-1}) \end{aligned} \tag{2}$$

Following Markov assumption, the current scene Gist ($G_t$) has all the necessary information for determining state and knowing the attended location. Thus $X_t$ is independent of all previous gists: $P(X_t|G_{1:t-1}) = P(X_t)$. Therefore, we can write: $P(X_t|G_{1:t}) = \mu P(G_t|X_t)P(X_t)$ with $P(X_t)$ as the prior distribution over eye positions.

**Case 2: Gist and previous saccade.** In the second step, we add the previous saccade locations to the formulation:

$$\begin{aligned} &P(X_t|G_{1:t}, X_{1:t-1}) \\ &= P(X_t|G_t, G_{1:t-1}, X_{1:t-1}) \\ &= \frac{P(G_t|X_t)P(X_t|G_{1:t-1}, X_{1:t-1})}{P(G_t|G_{1:t-1}, X_{1:t-1})} \\ &= \mu_1 P(G_t|X_t)P(X_t|G_{1:t-1}, X_{1:t-1}) \\ &= \mu_1\mu_2 P(G_t|X_t)P(X_{t-1}|X_t)P(X_t|G_{1:t-1}, X_{1:t-2}) \end{aligned} \tag{3}$$

where $\mu_1$ is equal to $P(G_t|G_{1:t-1}, X_{1:t-1})^{-1}$ and $\mu_2$ is $P(X_{t-1}|G_{1:t-1}, X_{1:t-2})^{-1}$. Again, considering Markov assumption and defining $\mu = \mu_1\mu_2$, we have: $P(X_t|G_{1:t}, X_{1:t-1}) = \mu P(G_t|X_t)P(X_{t-1}|X_t)P(X_t)$.

**Case 3: Gist, previous saccade, and motor actions.** Finally, we combine all evidences in our Bayesian model. Following the steps in case 2 and simplifying we reach to:

$$\begin{aligned} &P(X_t|G_{1:t}, X_{1:t-1}, A_{1:t-1}^{j=1:n}) \\ &= \mu P(G_t|X_t)P(X_{t-1}|X_t)P(X_t) \times \Pi_{j=1}^{n} P(A_{t-1}^{j}|X_t) \end{aligned} \tag{4}$$

Above formula assumes that actions are independent of each other given attended location (i.e., $A^k \perp A^l|X$). An important point here is whether actions influence saccades or vice-versa. In the real world the interaction works both ways: for some situations/tasks, saccades lead actions, however, sometimes actions can also lead eye movements. Here to be on the safe side, we did not use the current action.

Computing (4) requires estimation of $P(G_t|X_t)$ and similarly others. This can be done in several ways using non-parametric probability density estimation techniques such as generalized Gaussian model, histogram estimation or kNNs. We adapted the Kernel Density Estimation (KDE) approach. One pdf is calculated for each spatial location:

$$P(G|x_i) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{G}_h(x - x_i) = \frac{1}{mh}\sum_{i=1}^{m} \mathcal{G}(\frac{x - x_i}{h}) \tag{5}$$

where $\mathcal{G}_h$ is a Gaussian kernel with smoothing parameter (sliding window or bandwidth) $h$ and $m$ is number of data points. We used a Matlab toolbox[3] for implementing KDE.

[3]Publicly available at: http://www.ics.uci.edu/~ihler/code/kde.html

## 2.2. Baseline Benchmark Models

To fully evaluate effectiveness of our model, we implemented the regression model put forward by Peters and Itti [6] as well as a nearest-neighbor classifier and two other brute-force yet powerful models.

**Linear Regression (REG).** This model does not take into account the temporal progress of a task and simply maps Gist of the scene to the eye position. Mathematically, the goal is to optimize the following objective function:

$$\begin{aligned} &\arg\min_{W} ||M \times W - X_{sacc}||^2 \\ &Subject\ \ to: W \geq 0. \end{aligned} \tag{6}$$

where $M$ indicates the matrix of feature vectors (only Gist feature is used in [6]) and $X$ is the matrix of eye positions (one fixation per frame). The least-squares solution of the above objective function is: $W = M^+ \times X$, where $M^+$ is the pseudo-inverse of matrix $M$ through SVD decomposition. In our experiments, we only take the largest eigenvalue of the SVD since this avoids numerical instability and results in higher accuracy. Given vector $E = (u, v)$ as the eye position over a $20 \times 15$ map (i.e., $w = 20, h = 15$) with $u \in [1, 20]$ and $v \in [1, 15]$, the gaze density map can then be represented by vector $X = [x_1, x_2, \ldots, x_{300}]$ with $x_i = 1$ for $i = u + (v - 1) \times 20$ and $x_i = 0$ otherwise. Finally, for each test frame, we compute feature vector $F$ and generate the predicted map $P = F \times W$ which is then reshaped to a $20 \times 15$ saliency map. The maximum of this map is used to direct spatial attention.

**k Nearest Neighbor Classifier (kNN)**. We also implemented a non-linear mapping from features to saccade locations. The attention map for a test frame is built from the distribution of fixations of its most similar frames in the training set. For each test frame, $k$ most similar frames (using the Euclidean distance) were found and then the predicted map was the weighted average of the fixation locations of these frames (i.e., $X^i = \frac{1}{k}\sum_{j=1}^{k} D(F^i, F^j)^{-1}X^j$ where $X^j$ is the fixation map of the $j-th$ most similar frame to frame $i$ which is weighted according to its similarity to frame $i$ in feature space (i.e., $D(F^i, F^j)^{-1}$). We chose parameter $k$ to be 10 which resulted in good performance over train data as well as reasonable speed.

In addition to the above, we also devised two brute-force yet powerful predictors. The first one is simply the average of all saccade positions which we call **Average Fixation Map (AFM)** during the time course of a task over all m training frames (i.e., $AFM = \frac{1}{m}\sum_{j=1}^{m} X^j$). In dynamic environments used in this paper, since frames are generated on the fly and there are few fixations per frame, aligning frames (contrary to movies) is not possible. If a method could dynamically predict eye movements on a frame-by-frame basis, then achieving a higher accuracy than AFM is possible. AFM map is also the solution of the regression with a constant input, and is the output of our Bayesian

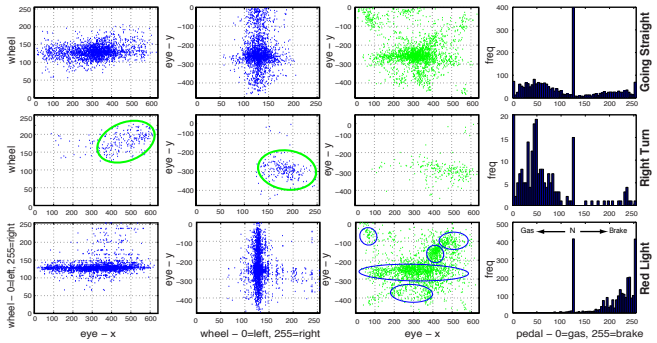



Figure 2. Correlation between actions and saccade positions. Rows indicate events (each frame was manually tagged based on its event). Columns from left to right include: $wheel$ vs. $eye-x$, $eye-y$ vs. $wheel$, saccade coordinates during the dame ($eye-x$ vs. $eye-y$), and frequency of pedal positions for DS game. Blue ellipses in the 3rd column indicate objects in the scene (see Fig. 1). Similar trends happen in the other games which eventually could help us in prediction of next saccade location.

model with one variable ($P(X)$ only). The second predictor is a **central Gaussian filter (Gauss)**. The rationale behind using this model is that humans tend to look at the center of the screen when game playing (center-bias or photographer-bias issue [31] by game design construction), therefore a central Gaussian blob may score well when datasets are centrally biased (See Figs. 3 and 6). Instead of using a fixed-size Gaussian for all games, we fitted a 2D Bivariate Gaussian to the fixation data of each game using ML algorithm:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\Big[\frac{-z}{2(1-\rho^2)}\Big] \quad (7)$$

where $z = \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}$ and $\rho$ is the correlation coefficient between $x$ and $y$ (i.e., $\rho = \frac{\sum_{xy}}{\sigma_x\sigma_y}$) where $\sum_{xy}$ is the covariance matrix.

## 3. Quantitative Results

Here we report results of our approach for predicting saccades (jumps in eye movements to bring the relevant object/location to the fovea)[4]. While we only process those frames in which a saccade happened, our method is easily applicable for predicting fixations (one for each frame).

### 3.1. Eye Tracking and Data Gathering

To test our models, we have collected a large amount of multi-modal data from subjects playing video games. We intend to share our data and accompany software to encourage follow-up research on modeling top-down attention.

Human subjects in the age range 20 to 30 played 5 video games. Subjects were students of anonymous university. Some subjects played more than one game. First, in a 5 min training session, aim and rules of the game as well as buttons of playing device were explained to the subject. Subjects were then asked to play the game to become familiar with the gaming environment. After training, in a test session, subjects played a different scenario of the game than during training (e.g., a different game level) without experimenter's intervention. They had different adventures in games from each other. Before the test session started, the eye tracker (ISCAN Inc. RK-464) was calibrated using 9 point calibration scheme. Subject's head was placed on a chin-rest at the distance of $130cm$ from the screen, yielding a visual field of $43^o \times 25^o$. Subject's right eye was recorded. Along with frames and fixations, subject's actions were also logged. A computer with Windows OS ran the PC games (frame rate $30Hz$), logged actions (frequency $62Hz$), and sent frames to a computer with Linux Mandriva OS that displayed and saved frames for later analysis. Another windows machine controlled the eye tracker camera and recorded fixations ($240Hz$). All computers communicated via a LAN network and their clocks were synchronized. Each data item had a time stamp which allowed us to align frame, action, and fixation data after recording.

**Stimuli.** To evaluate the power of our model, we applied it to 5 games with different task algorithms and visual renderings. For some games, scenes change considerably but for some others background scene is nearly constant making gist features less variable and informative.

Two of the games are driving games. The first one, *3D Driving School (DS)* is a driving emulator with simulated traffic conditions. Players must follow the route and European traffic rules defined by the game. An instructor will tell the players where to go by a text in a semi-translucent box above the screen and/or a small arrow on the top-left corner. Players use automatic transmission to drive around the entire course. This game has only dashboard view, an inside view from the driver-side towards the road. The second driving game, *18 Wheels of Steel (WS)* is a semi/truck simulator. In this game, players control a big rig to a specific destination, to retrieve money rewards for delivering a trailer. Players must drive carefully as the truck cannot accelerate/brake suddenly due to its mass. In this game, players were told to always make a left turn since there is no explicit instruction on the screen telling where to go. Players also used first-person/bumper view. Correlations between fixation patterns and driving events were found that can help detecting driver behavior's and intention (Fig. 2). Fig. 3 shows the average fixation map for DS game and it's corre-

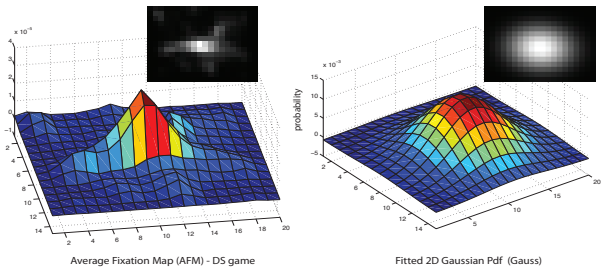



Figure 3. Average fixation map for the DS game and it's corresponding learned Gaussian map $\mu = [8.75\ 10.5]$ and $\sum = [7.85\ 0.52; 0.52\ 14.5]$.

[4]Thresholds to detect saccades were set to a velocity of $20^\circ/s$ and an amplitude threshold of $2^\circ/s$.

| Game | # Sacc. | # Subj | Dur. | # Frames | Size | Action |
|---|---|---|---|---|---|---|
| **DS** | 6382 | 10 | 10 min | 180K | 110 | J |
| **WS** | 4849 | 10 | 10 ” | 180K | 110 | J |
| **SM** | 1482 | 5 | 5 ” | 45K | 26 | J |
| **BS** | 1763 | 5 | 5 ” | 45K | 26 | M |
| **TG** | 4602 | 12 | $\sim 4.5$ ” | 99K | 57 | N/A |

Table 1. Summary statistics of our data including overall number of saccades, subjects, durations per subject, frames, sizes in GB, and action types (J indicates joystick and M stands for mouse).

sponding fitted Gaussian model.

The third game, *Super Mario Bros (SM)*, is a classic 2D-side-scrolling action game. Players control Mario to a flagpole to finish the level. Mario grows bigger if it consumes a mushroom and can shoot fireballs if it consumes a flower. There are various enemies that can be killed by stomping on them or shooting fireballs. In this game, players were expected not to take any means of shortcut such as running on ceiling, teleport pipes, or warp points. Actions in this game are $(x, y)$ position of joystick ([0, 255] for left/right, up/bottom) and status of 3 binary buttons including *Start, Jump,* and *Fire/Run*.

The fourth game called *Burger Shop (BS)* is a 2D time-management game. Under time pressure, players serve customers who order food items such as burgers and fries that must be assembled from a conveyor belt that brings ingredients. The game ends when all customers are served. For this game, actions include mouse $(x, y)$ position as well as status of the mouse buttons (i.e., *Left, Middle,* and *Right*).

The fifth game, *Top Gun (TG)*, is a flight-combat simulator. Players control a jet-fighter plane that can lock targets and shoot missiles, use afterburners to speed up, and do air maneuvers. The main objective of the game is to completely destroy all targets on air and on the ground. Players use first-person view in this stimuli. Currently, we do not have motor actions for this game.

Table. 1 shows summary statistics of video game data.

### 3.2. Evaluation Metrics

To quantify how well model predictions matched observers' actual eye positions, we used two metrics:

**Normalized Scanpath Saliency (NSS).** NSS [34] is defined as the response value at the human eye position $(x_h, y_h)$ in a model's predicted gaze density map that has been normalized to have zero mean and unit standard deviation: $NSS(t) = \frac{1}{\sigma_{s(x)}}\big(s(x(t)) - \mu_{s(t)}\big)$ for frame at time $t$. An NSS value of unity indicates the subject's eye position falls on a region whose predicted density is one standard deviation above average. Meanwhile, an NSS value of zero or

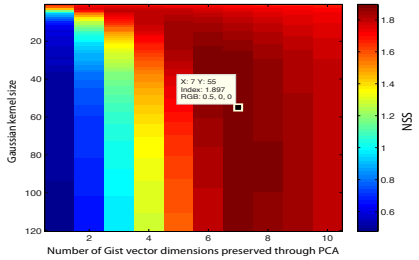



Figure 4. Grid search for best parameters (KDE kernel width and PCA dimensions of the Gist vector; Sec. 2) for DS game over train data.

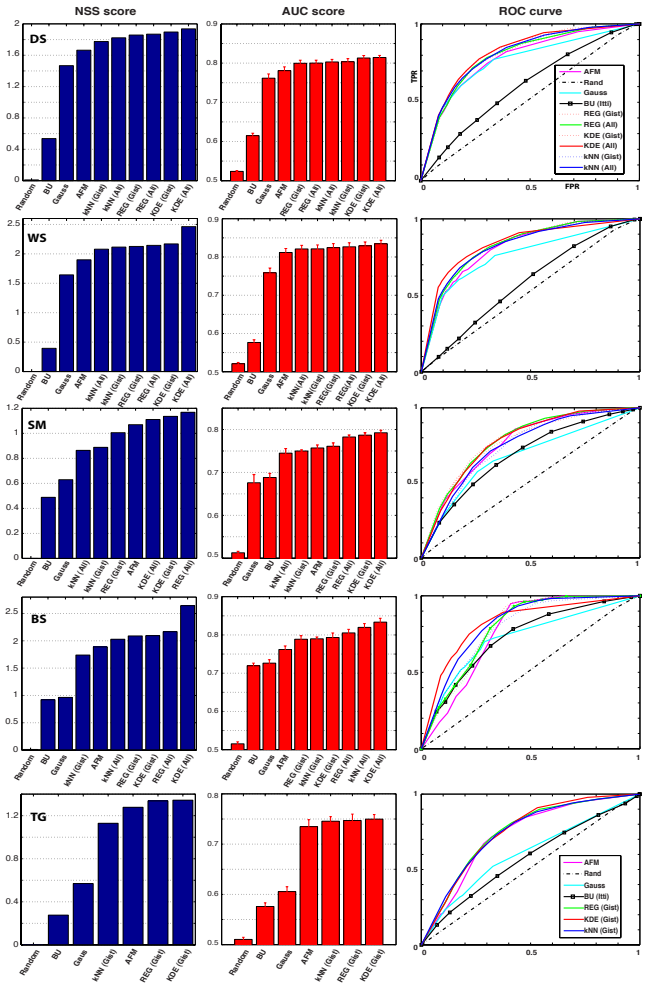



Figure 5. Prediction accuracy of our KDE model, Itti *et al.* [7], classifiers also implemented here, as well as brute-force predictors (AFM and Gaussian) for 5 video games using NSS and AUC (ROC) scores. KDE model with all features, KDE (All), results in the best performance in all cases, KDE with only Gist feature outperforms the other compared models.

lower means that the model performs no better than picking a random position on the map.

**Area Under the Curve (AUC).** Here, a model's saliency map is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated. Human fixations are used as ground truth. By varying the threshold, the Receiver Operating Characteristic (ROC) curve is drawn as the false positive rate vs. true positive rate, and the area under this curve indicates how well the saliency map predicts actual human eye fixations [14]. Perfect prediction corresponds to a score of 1.

**Results.** In the first experiment, we trained the model over each separate game. Each game segment has a variable number of saccades for each subject. Training was done over saccades of $K - 1$ subjects and tested over saccades of the remaining test subject. In each training phase, the best kernel width and PCA dimensions of gist vector (see Sec. 2) were found using grid search. Fig. 4 shows

| Game | ICL [17] | SDSR [20] | GBVS [24] | AIM [14] | SUN [19] | Gauss [31] | AFM | KDE(C-1) | KDE(C-2) | KDE(C-3) |
|---|---|---|---|---|---|---|---|---|---|---|
| **DS** | 0.57 | 0.54 | 0.73 | 0.62 | 0.658 | 0.76 | 0.78 | **0.82** | **0.82** | **0.82** |
| | 0.19 | 0.05 | 0.948 | 0.54 | 0.30 | 1.47 | 1.66 | **1.9** | **1.91** | **1.95** |
| **WS** | 0.52 | 0.41 | 0.73 | 0.55 | 0.51 | 0.76 | 0.81 | **0.83** | **0.83** | **0.84** |
| | 0.27 | -0.2 | 1.25 | 0.66 | 0.19 | 1.64 | 1.9 | **2.18** | **2.21** | **2.46** |
| **SM** | 0.61 | 0.69 | 0.72 | 0.67 | 0.62 | 0.67 | 0.75 | **0.78** | **0.79** | **0.79** |
| | 0.59 | 0.74 | **1.21** | 0.77 | 0. 33 | 0.62 | 1.07 | **1.13** | **1.21** | 1.11 |
| **BS** | 0.72 | 0.61 | 0.73 | 0.69 | 0.72 | 0.72 | 0.76 | **0.79** | **0.81** | **0.84** |
| | 1.04 | 0.54 | 1.1 | 0.80 | 1.2 | 0.96 | 1.89 | **2.1** | **2.2** | **2.7** |
| **TG** | 0.62 | 0.5 | 0.622 | 0.6 | 0.6 | 0.6 | **0.73** | **0.75** | **0.75** | - |
| | 0.58 | 0.01 | 0.55 | 0.51 | 0.29 | 0.57 | **1.28** | **1.36** | **1.34** | - |

Table 2. AUC(1st rows) and NSS scores(2nd rows) of 5 state-of-the-art models and ours over our data. Numbers in bold show best two models in each row. In almost all cases, while other models fall below Gaussian and AFM models, KDE (All) scores the best. In some cases, regression and KNN may score the best (cf. Fig. 5). C-x stands for Case x (See Sec. 2.1).

an example of best parameters over one training session of the DS game. Fig. 5 shows NSS and AUC scores, as well as ROC curves for baseline models and all variants of our model for each individual game. Over all games, KDE with all features (case 3) resulted in the best performance followed by case 2: KDE (Gist + Prev. sacc). KDE with only Gist feature outperformed classifiers with Gist, which indicates advantage of the KDE approach for using this feature compared with regression [6]. Random predictor (a random value for each location) has zero NSS and AUC near 0.5. AFM predictor achieved higher scores than BU [7] and Gaussian models over all games, indicating that eye movements were likely mostly guided top-down and BU influences were weak. AFM outperformed classifiers over the SM game, indicating that Gist is not a good predictor for this game; but when we added previous saccade position and actions, classifiers and KDE performed the best. Using action features alone in the kNN classifier resulted in NSSs of 1.41 and 1.80 for the DS and WS games, respectively higher than Gaussian and close to AFM of each game.

Table 2, shows accuracy of 5 state-of-the-art bottom-up saliency models, Gaussian, and AFM. In previous research, these models have achieved the highest scores over eye movement datasets for free-viewing task. Here, almost all of these models perform worse than AFM, while our approaches (KDE (All) and KDE (Gist)) perform higher with a large margin. This again indicates that, while bottom-up saliency models fail to account for eye fixations in our tasks which have a strong top-down component, our new models are able to capture a large amount of task-driven saccades.

In the second experiment, we trained the KDE models over one of two driving games and tested it on the other to assess the generalization power of our approach over different tasks. As Table 3 shows, training on a similar game

| **Train on** | DS | WS |
|---|---|---|
| **Test on** | WS | DS |
| **AFM** | 0.80 (1.74) | 0.75 (1.51) |
| **KDE (Gist)** | 0.80 (1.64) | 0.74 (1.40) |
| **KDE (All)** | 0.79 (1.62) | 0.73 (1.51) |

Table 3. Confusion matrix of training models on one driving game and applying it to the other one using AUC and NSS(parenthesis).

| | Gist [10] | | HOG [30] | |
|---|---|---|---|---|
| Game | kNN | REG | kNN | REG |
| **DS** | 0.80 (1.77) | 0.8 (1.86) | 0.81 (1.88) | **0.81 (2.05)** |
| **SM** | 0.75 (0.88) | 0.76 (1.01) | 0.74 (0.97) | **0.79 (1.23)** |

Table 4. Comparing AUC and NSS (in parenthesis) of Gist model of Siagian *et al.* [10] and HOG features for saccade prediction using kNN and regression classifiers for 3D Driving School and Super Mario games. Dimensionality of Gist vector is 714 and dimensionality of HOG is 4800. Only for REG (HOG) dimensionality of HOG is reduced to 95% of its variance which preserved about 900 D for DS and 500 for Mario game.

results in higher accuracy than random, and close to performance of Gaussian and AFM predictors of each game shown in Table 2. Applying the AFM of games to each other resulted in higher accuracy than the KDE models, probably because one constant in both games is that subjects look at the center. Since actions and sequence of fixations are specific for each game, adding them slightly drops the performance (KDE (Gist) vs. KDE (All)).

In the third experiment, we aimed to compare the power of HOG features [30] and the Gist features of Siagian *et al.* [10]. The notion behind using HOG features is that they encode rich structural information from the entire scene and have been very successful in object detection. Table 4 shows the performance of kNN and regression classifiers over DS and SM games. HOG features were better descriptors of the scene and conveyed more information regarding saccade locations over both games and using both classifiers. However, because calculating 8 orientation channels in HOG makes it slower than gist in [10] (about 2 times) which uses 4, here we performed experiments using the second one. HOG also generates high dimensional feature vectors which makes it hard to store and work with.

Figure 6 shows sample frames of video games with corresponding saliency maps from models. Predicted maps by our models show dense activity at task relevant locations thereby narrowing attention and leading to higher NSS and AUC scores. These maps change per frame as opposed to the static AFM and Gaussian models.

## 4. Discussion and Future Work

We proposed a unified Bayesian approach that is applicable to a large class of everyday tasks where global scene knowledge, the sequence of fixated locations, and actions, constrain future eye fixations. In addition to the above-mentioned factors, there might be other general features influencing task-driven attention. Our framework allows easy incorporation of those features for saccade prediction.

An important application of our model is quantitative analysis of differences among populations of subjects (e.g., young vs. elderly or novices vs. experts) in complex tasks such as driving. It can also be useful for assistant technologies for demanding tasks, human computer interaction, context aware systems, and health care.

Although employed features convey information regarding the next saccade, it is still possible to gain higher performance by knowing more about the scene. For instance, by

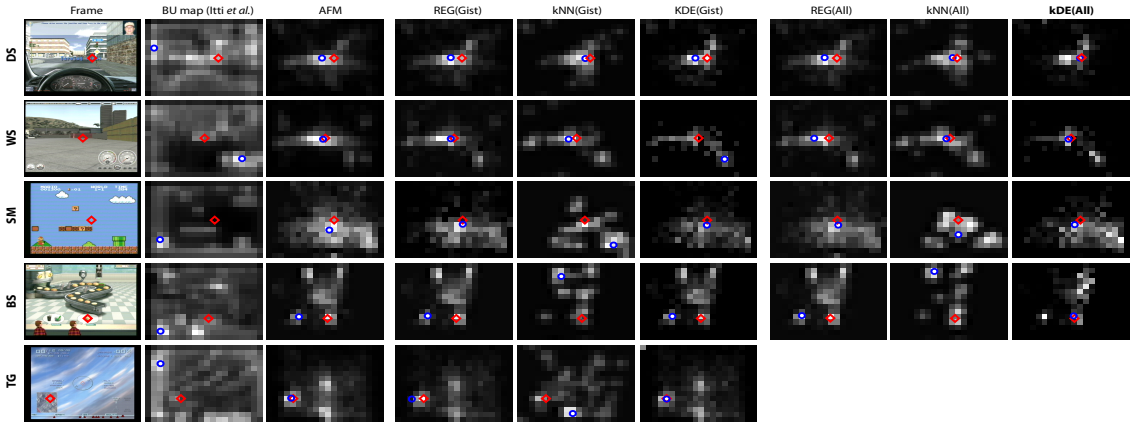



Figure 6. Sample frames of video games and corresponding predicted maps of models. Red diamond indicates the human fixation and blue circles is the maximum point of each map. Smaller distance hence means better prediction. Currently we don't have action data for TG game.

calculating the number or state of task-related objects. Such approach, however, has the drawback that for each task, relevant variables and interactions among them should be defined, thus limiting its generalization. We are now investigating the role of local context ($P(O|L_v)$ in (1)) in modulation of top-down attention. Instead of predicting fixation locations, it may be more efficient to bias the visual system toward features of a relevant object within a global context. While the exact fixated location at nearly the same gist may change based on recent history of saccades and actions, looking for a given object rather than a given location may exhibit stronger invariance. Also, extraction and addition of subjective factors such as fatigue, preference, and experience into our model would be an interesting next step.

*Supported by the National Science Foundation (grant number BCS-0827764), and the Army Research Office (W911NF-08-1-0360 and W911NF-11-1-0046), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.*

## References

[1] A. Yarbus. Eye movements during perception of complex objects. L. Riggs, editor, Eye Movements and Vision, 1967. 2

[2] M. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 2001. 2

[3] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cog. Sci.*, 9(4):188-193. 2005. 2

[4] D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *J. of Cog. Neurosci.*, 7(1):66-80, 1995. 2, 3

[5] M.F. Land and D.N. Lee. Where we look when we steer. *Nature*, 1994. 2

[6] R.J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *CVPR*, 2007. 2, 4, 7

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998. 1, 2, 3, 6, 7

[8] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2): 205-231, 2005. 2

[9] N. Sprague and D. Ballard. Eye Movements for reward maximization. *NIPS*, 2003. 2

[10] C. Siagian and L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention. *PAMI*, 29(2):300-312, 2007. 2, 3, 7

[11] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: a combined source model of eye guidance. *Visual Cognition*, 2009. 1, 2

[12] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psych.*, 12:97-136, 1980. 1

[13] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 1985. 1

[14] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005. 1, 2, 6, 7

[15] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences and applications to visual recognition. *IEEE PAMI*, 2009. 1

[16] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *CVPR*, 2007. 1

[17] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *NIPS*, 2008. 1, 7

[18] A. Torralba, A. Oliva, M. Castelhano and J. M. Henderson. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 2006. 1, 2, 3

[19] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, SUN: A Bayesian framework for saliency using natural statistics. *J. Vis.*, 2008. 1, 2, 7

[20] H. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 2009. 7

[21] T. Avraham, M. Lindenbaum. ESaliency: Meaningful attention using stochastic image modeling. *PAMI*, 2010. 1

[22] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed.*CVPR*,2006. 1, 2

[23] S. Frintrop. VOCUS: A visual attention system for object detection and goal-directed search. Springer 2006. 2

[24] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *NIPS*, 2006. 7

[25] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *J. of Vision*, 8:1-17, 2008. 1

[26] W. Einhauser, U. Rutishauser, and C. Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *J. of Vision*, 2008. 1

[27] J. Henderson. Regarding scenes. *Current directions in psychological science*, 16:219-227, 2007. 1

[28] F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *PNAS*, 2002. 1

[29] S. Ullman. Visual routines. *Cognition*, 18:97-157, 1984. 2

[30] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005. 3, 7

[31] B.W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. *J. of Vision*. 14(7):1-17, 2007. 5, 7

[32] J. Hays, A.A. Efros. Scene completion using millions of photographs. *SIGGRAPH*, 2007. 2

[33] L. Renniger and J. Malik. When is scene identication just texture recognition? *Vision Research*, 2004. 3

[34] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Res.*, 45, 2005. 6

[35] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? *CVPR*, 2010. 1

[36] S. Vijayanarasimhan and A. Kapoor. Visual recognition and detection under bounded computational resources. *CVPR*, 2010. 1

[37] H.W. Kang, Y. Matsushita, X. Tang, and X.Q. Chen. Space-time video montage. *CVPR*, 2006. 1

[38] L. Wolf, M. Guttmann, and D. Cohen-Or. Content-driven video retargeting. *ICCV*, 2007. 1

# State-of-the-art in Visual Attention Modeling

Ali Borji, *Member, IEEE,* and Laurent Itti, *Member, IEEE*

**Abstract**—Modeling visual attention — particularly stimulus-driven, saliency-based attention — has been a very active research area over the past 25 years. Many different models of attention are now available, which aside from lending theoretical contributions to other fields, have demonstrated successful applications in computer vision, mobile robotics, and cognitive systems. Here we review, from a computational perspective, the basic concepts of attention implemented in these models. We present a taxonomy of nearly 65 models, which provides a critical comparison of approaches, their capabilities, and shortcomings. In particular, thirteen criteria derived from behavioral and computational studies are formulated for qualitative comparison of attention models. Furthermore, we address several challenging issues with models, including biological plausibility of the computations, correlation with eye movement datasets, bottom-up and top-down dissociation, and constructing meaningful performance measures. Finally, we highlight current research trends in attention modeling and provide insights for future.

**Index Terms**—Visual attention, bottom-up attention, top-down attention, saliency, eye movements, regions of interest, gaze control, scene interpretation, visual search, gist.

## 1 Introduction

A RICH stream of visual data ($10^8 - 10^9$ bits) enters our eyes every second [1][2]. Processing this data in real-time is an extremely daunting task without the help of clever mechanisms to reduce the amount of erroneous visual data. High-level cognitive and complex processes such as object recognition or scene interpretation rely on data that has been transformed in such a way to be tractable. The mechanism this paper will discuss is referred to as visual attention - and at its core lies an idea of a selection mechanism and a notion of relevance. In humans, attention is facilitated by a retina that has evolved a high-resolution central fovea and a low-resolution periphery. While visual attention guides this anatomical structure to important parts of the scene to gather more detailed information, the main question is on the computational mechanisms underlying this guidance.

In recent decades, many facets of science have been aimed towards answering this question. Psychologists have studied behavioral correlates of visual attention such as change blindness [3][4], inattentional blindness [5], and attentional blink [6]. Neurophysiologists have shown how neurons accommodate themselves to better represent objects of interest [27][28]. Computational neuroscientists have built realistic neural network models to simulate and explain attentional behaviors (e.g., [29][30]). Inspired by these studies, robotists and computer vision scientists have tried to tackle the inherent problem of computational complexity to build systems capable of working in real-time (e.g., [14][15]). Although there are many models available now in the research areas mentioned above, here we limit ourselves to models that can compute saliency maps (please see next section for definitions) from any image or video input. For a review on computational models of visual attention in general, including biased competition [10], selective tuning [15], normalization models of attention [181], and many others, please refer to [8]. Reviews of attention models from psychological, neurobiological, and computational perspectives can be found in [9][77][10][12][202][204][224]. Fig. 1 shows a taxonomy of attentional studies and highlights our scope in this review.

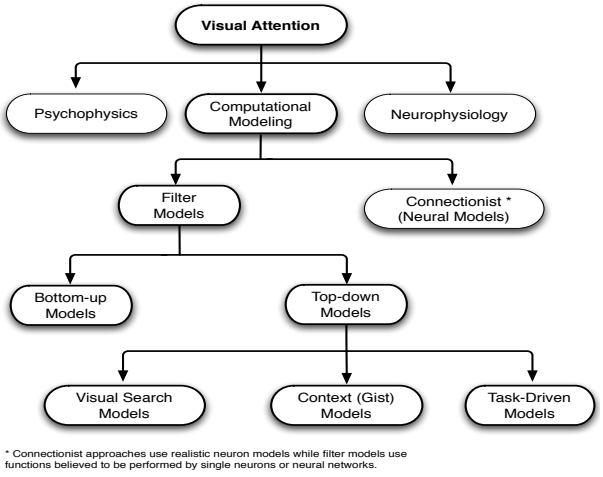



Fig. 1. Taxonomy of visual attention studies. Ellipses with solid borders illustrate our scope in this paper.

### 1.1 Definitions

While the terms attention, saliency, and gaze are often used interchangeably, each has a more subtle definition that allows their delineation.

Attention is a general concept covering all factors that influence selection mechanisms, whether they be scene-driven bottom-up (BU) or expectation-driven top-down (TD).

Saliency intuitively characterizes some parts of a scene — which could be objects or regions — that appear to an observer to stand out relative to their neighboring parts. The term "salient" is often considered in the context of bottom-up computations [18][14].

Gaze, a coordinated motion of the eyes and head, has often been used as a proxy for attention in natural behavior

• *Authors are with the Department of Computer Science, University of Southern California (USC), Los Angeles, CA, 90089. E-mail: {borji,itti}@usc.edu*

• *Manuscript received November 2010.*

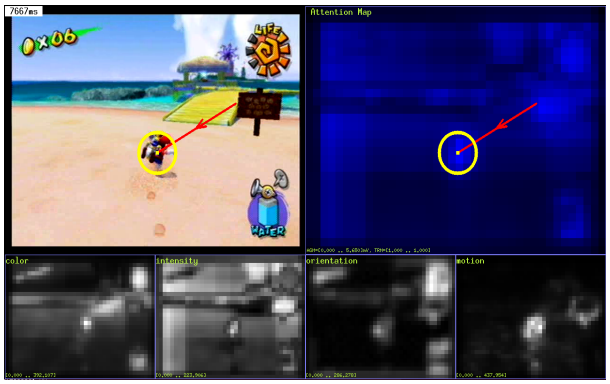



Fig. 2. Neuromorphic Vision C++ Toolkit (iNVT) developed at iLab, USC, http://ilab.usc.edu/toolkit/. A saccade is targeted to the location that is different from its surroundings in several features. In this frame from a video, attention is strongly driven by motion saliency.

(see [99]). For instance, a human or a robot has to interact with surrounding objects and control the gaze to perform a task while moving in the environment. In this sense, gaze control engages vision, action, and attention simultaneously to perform sensorimotor coordination necessary for the required behavior (e.g., reaching and grasping).

### 1.2 Origins

The basis of many attention models dates back to Treisman & Gelade's [81] "Feature Integration Theory" where they stated which visual features are important and how they are combined to direct human attention over pop-out and conjunction search tasks. Koch and Ullman [18] then proposed a feed-forward model to combine these features and introduced the concept of a saliency map which is a topographic map that represents conspicuousness of scene locations. They also introduced a winner-take-all neural network that selects the most salient location and employs an inhibition of return mechanism to allow the focus of attention to shift to the next most salient location. Several systems were then created implementing related models which could process digital images [15][16][17]. The first complete implementation and verification of the Koch & Ullman model was proposed by Itti *et al.* [14] (see Fig. 2) and was applied to synthetic as well as natural scenes. Since then, there has been increasing interest in the field. Various approaches with different assumptions for attention modeling have been proposed and have been evaluated against different datasets. In the following sections, we present a unified conceptual framework in which we describe the advantages and disadvantages of each model against one another. We give the reader insight into the current state of the art in attention modeling and identify open problems and issues still facing researchers.

The main concerns in modeling attention are how, when, and why we select behaviorally-relevant image regions. Due to these factors, several definitions and computational perspectives are available. A general approach is to take inspiration from the anatomy and functionality of the early human visual system, which is highly evolved to solve these problems (e.g., [14][15][16][191]). Alternatively, some studies have hypothesized what function visual attention may serve and have formulated it in a computational framework. For instance, it has been claimed that visual attention is attracted to the most informative [144], the most surprising scene regions [145], or those regions that maximize reward regarding a task [109].

### 1.3 Empirical Foundations

Attentional models have commonly been validated against eye movements of human observers. Eye movements convey important information regarding cognitive processes such as reading, visual search, and scene perception. As such, they often are treated as a proxy for shifts of attention. For instance, in scene perception and visual search, when the stimulus is more cluttered, fixations become longer and saccades become shorter [19]. The difficulty of the task (e.g., reading for comprehension versus reading for gist, or searching for a person in a scene versus looking at the scene for a memory test) obviously influences eye movement behavior [19]. Although both attention and eye movement prediction models are often validated against eye data, there are slight differences in scope, approaches, stimuli, and level of detail. Models for eye movement prediction (saccade programming) try to understand mathematical and theoretical underpinnings of attention. Some examples include search processes (e.g., optimal search theory [20]), information maximization models [21], Mr. Chips: an ideal-observer model of reading [25], EMMA (Eye Movements and Movement of Attention) model [139], HMM model for controlling eye movements [26], and constrained random walk model [175]). To that end, they usually use simple controlled stimuli, while on the other hand, attention models utilize a combination of heuristics, cognitive and neural evidence, and tools from machine learning and computer vision to explain eye movements in both simple and complex scenes. Attention models are also often concerned with practical applicability. Reviewing all movement prediction models is beyond the scope of this paper. The interested reader is referred to [22][23][127] for eye movement studies and [24] for a breadth-first survey of eye tracking applications.

Note that eye movements do not always tell the whole story and there are other metrics which can be used for model evaluation. For example, accuracy in correctly reporting a change in an image (i.e., search-blindness [5]), or predicting what attention grabbing items one will remember, show important aspects of attention which are missed by sole analysis of eye movements. Many attention models in visual search have also been tested by accurately estimating reaction times (RT) (e.g., RT/setsize slopes in pop-out and conjunction search tasks [224][191]).

### 1.4 Applications

In this paper, we focus on describing the attention models themselves. There are, however, many technological applications of these models which have been developed over the years and which have further increased interest in attention modeling. We organize the applications of attention modeling into three categories: vision and graphics, robotics, and those in other areas as shown in Fig. 3.

### 1.5 Statement and Organization

Attention is difficult to define formally in a way that is universally agreed upon. However, from a computational

| Category | Application | References |
|---|---|---|
| Computer Vision and Graphics | Image segmentation | Mishra and Aloimonos, 2009, Maki et al., 2000 |
| | Image quality assessment | Ma and Zhang, 2008, Ninassi et al., 2007 |
| | Image matching | Walther et al., 2006, Siagian and Itti, 2009, Frintrop and Jensfelt, 2008 |
| | Image rendering | DeCarlo and Santella, 2002 |
| | Image and video compression | Ouerhani et al., 2003, Itti, 2004, Guo and Zhang, 2010. |
| | Image thumbnailing | Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003 |
| | Image super-resolution | Jacobson et al, 2010 |
| | Image re-targeting (thumbnailing) | Setlur et al., 2005, Chamaret et al., 2008, Goferman et al., 2010, Achanta et al., 2009, Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003 |
| | Image superresolution | Sadaka and Karam, 2009 |
| | Video summarization | Marat et al., 2007, Ma et al., 2005 |
| | Scene classification | Siagian and Itti, 2009 |
| | Object detection | Frintrop, 2006, Navalpakkam and Itti, 2006, Fritz et al., 2005, Butko and Movellan, 2009, Viola and Jones, 2004, Ehinger et al., 2009. |
| | Salient object detection | Liu et al, 2007, Goferman et al. 2010, Achanta et al., 2009, Rosin, 2009. |
| | Object recognition | Salah et al., 2002, Walther et al., 2006 and 2007, Frintrop, 2006, Mitri et al., 2005, Gao and Vasconcelos, 2004 and 2009, Han and Vasconcelos 2010, Paletta et al., 2005. |
| | Visual tracking | Mahadevan and Vasconcelos, 2009, Frintrop, 2010 |
| | Dynamic lighting | Seif El-Nasr, 2009 |
| | Video shot detection | Boccignone et al., 2005 |
| | Interest point detection | Kadir and Brady, 2001, Kienzle et al., 2007. |
| | Automatic collage creation | Goferman et al., 2010, Wang et al., 2006. |
| | Face segmentation and tracking | Li and Ngan, 2008 |
| Robotics | Active vision | Mertsching et al., 1999, Vijaykumar et al., 2001, Dankers, 2007, Borji et al., 2010 |
| | Robot Localization | Siagian and Itti, 2009, Ouerhani et al., 2005 |
| | Robot Navigation | Baluja and Pomerlau, 1997, Scheier and Egner, 1997 |
| | Human-robot interaction | Breazeal, 1999, Heidemann et al., 2004, Belardinelli, 2008, Nagai, 2009, Muhl, 2007 |
| | Synthetic vision for simulated actors | Courty and Marchand, 2003 |
| Others | Advertising | Rosenholtz et al. 2011, Liu et al., 2008 |
| | Finding tumors in mammograms | Hong and Brady, 2003 |
| | Retinal prostheses | Parick et al., 2010 |

Fig. 3. Some applications of visual attention modeling.

standpoint, many models of visual attention (at least those tested against first few seconds of eye movements in free-viewing) can be unified under the following general problem statement. Assume $K$ subjects have viewed a set of $N$ images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^{N}$. Let $\boldsymbol{L}_i^k = \{\boldsymbol{p}_{ij}^k, t_{ij}^k\}_{j=1}^{n_i^k}$ be the vector of eye fixations (saccades) $\boldsymbol{p}_{ij}^k = (x_{ij}^k, y_{ij}^k)$ and their corresponding occurrence time $t_{ij}^k$ for the $k$-th subject over image $\boldsymbol{I}_i$. Let the number of fixations of this subject over $i$-th image be $n_i^k$. The goal of attention modeling is to find a function (stimuli-saliency mapping) $f \in \mathcal{F}$ which minimizes the error on eye fixation prediction, i.e, $\sum_{k=1}^{K}\sum_{i=1}^{N} m(f(\boldsymbol{I}_i^k), \boldsymbol{L}_i^k)$, where $m \in \mathcal{M}$ is a distance measure (defined in section 2.7). An important point here is that the above definition better suits bottom-up models of overt visual attention, and may not necessarily cover some other aspects of visual attention (e.g., covert attention or top-down factors) that cannot be explained by eye movements.

Here we present a systematic review of major attention models that we apply to arbitrary images. In section 2, we first introduce several factors to categorize these models. In section 3, we then summarize and classify attention models according to these factors. Limitations and issues in attention modeling are then discussed in section 4 and are followed by conclusions in section 5.

## 2 Categorization Factors

We start by introducing 13 factors ($\mathtt{f}_{1..13}$) that will be used later for categorization of attention models. These factors have their roots in behavioral and computational studies of attention. Some factors describe models ($\mathtt{f}_{1,2,3}$, $\mathtt{f}_{8..11}$), others ($\mathtt{f}_{4..7}$, $\mathtt{f}_{12,13}$) are not directly related, but are just as important as they determine the scope of applicability of different models.

### 2.1 Bottom-up vs. Top-down Models

A major distinction among models is whether they rely on bottom-up influences ($\mathtt{f}_1$), top-down influences ($\mathtt{f}_2$), or a combination of both.

Bottom-up cues are mainly based on characteristics of a visual scene (stimulus-driven)[75], whereas top-down cues (goal-driven) are determined by cognitive phenomena like knowledge, expectations, reward, and current goals.

Regions of interest that attract our attention in a bottom-up manner must be sufficiently distinctive with respect to surrounding features. This attentional mechanism is also called exogenous, automatic, reflexive, or peripherally cued [78]. Bottom-up attention is fast, involuntary, and most likely feed-forward. A prototypical example of bottom-up attention is looking at a scene with only one horizontal bar among several vertical bars where attention is immediately drawn to the horizontal bar [81]. While many models fall in this category, they can only explain a small fraction of eye movements since the majority of fixations are driven by task [177].

On the other hand, top-down attention is slow, task-driven, voluntary, and closed-loop [77]. One of the most famous examples of top-down attention guidance is from Yarbus in 1967 [79], who showed that eye movements depend on the current task with the following experiment: subjects were asked to watch the same scene (a room with a family and an unexpected visitor entering the room) under different conditions (questions) such as "estimate the material circumstances of the family", "what are the ages of the people?", or simply to freely examine the scene. Eye movements differed considerably for each of these cases.

Models have explored three major sources of top-down influences in response to this question: How do we decide where to look?. Some models address visual search in which attention is drawn toward features of a target object we are looking for. Some other models investigate the role of scene context or gist to constrain locations that we look at. In some cases, it is hard to precisely say where or what we are looking at since a complex task governs eye fixations, for example in driving. While in principle, task demands on attention subsumes the other two factors, in practice models have been focusing on each of them separately. Scene layout has also been proposed as a source of top-down attention [80][93] and is here considered together with scene context.

1) **Object Features**. There is a considerable amount of evidence for target-driven attentional guidance in real-world search tasks [84][85][23][83]. In classical search tasks, target features are a ubiquitous source of attention guidance [81][82][83]. Consider a search over simple search arrays in which the target is a red item: attention is rapidly directed toward the red item in the scene. Compare this with a more complex target object, such as a pedestrian in a natural scene, where although it is difficult to define the target, there are still some features (e.g., upright form, round head, and straight body) to direct visual attention [87].

The guided search theory [82] proposes that attention can be biased toward targets of interest by modulating the relative gains through which different features contribute to attention. To return to our prior example, when looking for a red object, a higher gain would be assigned to red color. Navalpakkam *et al.* [51] derived the optimal integration of cues (channels of the BU saliency model [14]) for detection of a target in terms of maximizing the signal-to-noise ratio of

the target versus background. In [50], a weighting function based on a measure of object uniqueness was applied to each map before summing up the maps for locating an object. Butko *et al.* [161] modeled object search based on the same principles of visual search as stated by Najemnik *et al.* [20] in a partially observable framework for face detection and tracking, but they did not apply it to explain eye fixations while searching for a face. Borji *et al.* [89] used evolutionary algorithms to search in a space of basic saliency model parameters for finding the target. Elazary and Itti [90] proposed a model where top-down attention can tune both the preferred feature (e.g., a particular hue) and the tuning width of feature detectors, giving rise to more flexible top-down modulation compared to simply adjusting the gains of fixed feature detectors. Last but not least are studies such as [147][215][141] that derive a measure of saliency from formulating search for a target object.

The aforementioned studies on the role of object features in visual search are closely related to object detection methods in computer vision. Some object detection approaches (e.g., Deformable Part Model by Felzenszwalb *et al.* [206] and the Attentional Cascade of Viola and Jones [220]) have high detection accuracy for several objects such as cars, persons, and faces. In contrast to cognitive models, such approaches are often purely computational. Research on how these two areas are related will likely yield mutual benefits for both.

2) **Scene Context**. Following a brief presentation of an image (∼80 ms or less), an observer is able to report essential characteristics of a scene [176][71]. This very rough representation of a scene, so called "gist", does not contain many details about individual objects but can provide sufficient information for coarse scene discrimination (e.g., indoor vs. outdoor). It is important to note that gist does not necessarily reveal the semantic category of a scene. Chun and Jiang [91] have shown that targets appearing in repeated configurations relative to some background (distractor) objects were detected more quickly [71]. Semantic associations among objects in a scene (e.g., a computer is often placed on top of a desk) or contextual cues have also been shown to play a significant role in the guidance of eye movements [199][84].

Several models for gist utilizing different types of low-level features have been presented. Oliva and Torralba [93], computed the magnitude spectrum of a Windowed Fourier Transform over non-overlapping windows in an image. They then applied principal component analysis (PCA) and independent component analysis (ICA) to reduce feature dimensions. Renninger and Malik [94] applied Gabor filters to an input image and then extracted 100 universal textons selected from a training set using K-means clustering. Their gist vector was a histogram of these universal textons. Siagian and Itti [95] used biological center-surround features from orientation, color, and intensity channels for modeling gist. Torralba [92] used wavelet decomposition tuned to 6 orientations and 4 scales. To extract gist, a vector is computed by averaging each filter output over a $4 \times 4$ grid. Similarly he applied PCA to the resultant 384D vectors to derive a 80D gist vector. For a comparison of gist models, please refer to [96][95].

Gist representations have become increasingly popular in computer vision since they provide rich global yet discriminative information useful for many applications such as search in the large-scale scene datasets available today [116], limiting the search to locations likely to contain an object of interest [92][87], scene completion [205], and modeling top-down attention [101][218]). It can thus be seen that research in this area has the potential to be very promising.

3) **Task Demands**. Task has a strong influence on deployment of attention [79]. It has been claimed that visual scenes are interpreted in a need-based manner to serve task demands [97]. Hayhoe *et al.* [99] showed that there is a strong relationship between visual cognition and eye movements when dealing with complex tasks. Subjects performing a visually-guided task were found to direct a majority of fixations toward task-relevant locations [99]. It is often possible to infer the algorithm a subject has in mind from the pattern of her eye movements. For example, in a "block-copying" task where subjects had to replicate an assemblage of elementary building blocks, the observers' algorithm for completing the task was revealed by patterns of eye-movements. Subjects first selected a target block in the model to verify the block's position, then fixated the workspace to place the new block in the corresponding location [216]. Other research has studied high-level accounts of gaze behavior in natural environments for tasks such as sandwich making, driving, playing cricket, and walking (see Henderson and Hollingworth [177], Rensink [178], Land and Hayhoe [135], and Bailensen and Yee [179]). Sodhi *et al.* [180] studied how distractors while driving such as adjusting the radio or answering a phone affect eye movements.

The prevailing view is that bottom-up and top-down attention are combined to direct our attentional behavior. An integration method should be able to explain when and how to attend to a top-down visual item or skip it for the sake of a bottom-up salient cue. Recently, [13] proposed a Bayesian approach that explains the optimal integration of reward as a top-down attentional cue, and contrast or orientation as a bottom-up cue in humans. Navalpakkam and Itti [80] proposed a cognitive model for task-driven attention constrained by the assumption that the algorithm for solving the task was already available. Peters and Itti [101] learned a top-down mapping from scene gist to eye fixations in video game playing. Integration was simply formulated as multiplication of BU and TD components.

### 2.2 Spatial vs. Spatio-temporal Models

In the real-world, we are faced with visual information that constantly changes due to egocentric movements or dynamics of the world. Visual selection is then dependent on both current scene saliency as well as the accumulated knowledge from previous time points. Therefore, an attention model should be able to capture scene regions that are important in a spatio-temporal manner.

To be detailed in section 3, almost all attention models include a spatial component. We can distinguish between two types of modeling temporal information in saliency modeling: 1) Some bottom-up models use the *motion* channel to capture human fixations drawn to moving stimuli [119]. More recently, several researchers have started modeling temporal effects on bottom-up saliency (e.g., [143][104][105]). 2) On the other hand, some models [109][218][26][25][102] aim to capture the spatio-temporal aspects of a task for example by learning sequences of attended objects or actions as the task progresses.

For instance, the Attention Gate Model (AGM) [183], emphasizes the temporal response properties of attention and quantitatively describes the order and timing for humans attending to sequential target stimuli. Previous information about images, eye fixations, image content at fixations, physical actions, as well as other sensory stimuli (e.g., auditory) can be exploited to predict the next eye movement. Adding a temporal dimension and the realism of natural interactive tasks brings a number of complications in predicting gaze targets within a computational model.

Suitable environments for modeling temporal aspects of visual attention are dynamic and interactive setups such as movies and games. Boiman and Irani [122] proposed an approach for irregularity detection from videos by comparing texture patches of actions with a learned dataset of irregular actions. Temporal information was limited to the stimulus level and did not include higher cognitive functions such as the sequence of items processed at the focus of attention or actions performed while playing the games. Some methods derive static and dynamic saliency maps and propose methods to fuse them (e.g., Jia Li *et al.* [133] and Marat *et al.* [49]). In [103], a spatio-temporal attention modeling approach for videos is presented by combining motion contrast derived from the homography between two images and spatial contrast calculated from color histograms. Virtual reality (VR) environments have also been used in [99][109][97]. Some other models dealing with the temporal dimension are [105][108][103]. We postpone the explanation of these approaches to section 3.

Factors $f_3$ indicates whether a model uses spatial only or spatio-temporal information for saliency estimation.

### 2.3 Overt vs. Covert attention

Attention can be differentiated based on its attribute as "overt" versus "covert". Overt attention is the process of directing the fovea towards a stimulus while covert attention is mentally focusing onto one of several possible sensory stimuli. An example of covert attention is staring at a person who is talking but being aware of visual space outside the central foveal vision. Another example is driving, where a driver keeps his eyes on the road while simultaneously covertly monitoring the status of signs and lights. The current belief is that covert attention is a mechanism for quickly scanning the field of view for an interesting location. This covert shift is linked to eye movement circuitry that sets up a saccade to that location (overt attention) [203]. However, this does not completely explain complex interactions between covert and overt attention. For instance, it is possible to attend to the right hand corner field of view and actively suppress eye movements to that location. Most of these models detect regions that attract eye fixations and few explain overt orientation of eyes along with head movements. Lack of computational frameworks for covert attention might be because behavioral mechanisms and functions of covert attention are still unknown. Further, it is not known yet how to measure covert attention.

Because of a great deal of overlap between overt and covert attention and since they are not exclusive concepts, saliency models could be considered as modeling both overt and covert mechanisms. However, in depth discussion of this topic goes beyond our scope and merits of this paper and demands special treatment elsewhere.

### 2.4 Space-based vs. Object-based Models

There is no unique agreement on the unit of attentional scale: Do we attend to spatial locations, to features, or to objects? The majority of psychophysical and neurobiological studies are about space-based attention (e.g., Posner's spatial cueing paradigm [98][111]). There is also strong evidence for feature-based attention (detecting an odd item in one feature dimension [81] or tuning curve adjustments of feature selective neurons [7]) and object-based attention (selectivity attending to one of two objects, e.g., face vs. vase illusion [112][113][84]). The current belief is that these theories are not mutually exclusive and visual attention can be deployed to each of these candidate units, implying there is no single unit of attention. Humans are capable of attending to multiple (between four and five) regions of interest simultaneously [114][115].

In the context of modeling, a majority of models are space-based (see Fig. 7). It is also viable to think that humans work and reason with objects (compared with rough pixel values) as main building blocks of top-down attentional effects [84]. Some object-based attentional models have previously been proposed, but they lack explanation for eye fixations (e.g., Sun and Fisher [117], Borji *et al.* [88]). This shortcoming makes verification of their plausibility difficult. For example, the limitation of the Sun and Fisher [117] model is the use of human segmentation of the images; it employs information that may not be available in the pre-attentive stage (before the objects in the image are recognized). Availability of object-tagged image and video datasets (e.g., LabelMe Image and Video [116][188]) has made conducting effective research in this direction possible. The link between object-based and space-based models remains to be addressed in the future. Feature-based models (e.g., [51][83]) adjust properties of some feature detectors in an attempt to make a target object more salient in a distracting background. Because of the close relationship between visual features and objects, in this paper we categorize feature-based models under object-based models as shown in Fig. 7.

The ninth factor $f_9$, indicates whether a model is space-based or object-based - meaning that it needs to work with objects instead of raw spatial locations.

### 2.5 Features

Traditionally, according to feature integration theory (FIT) and behavioral studies [81][82][118], three features have been used in computational models of attention: intensity (or intensity contrast, or luminance contrast), color, and orientation. Intensity is usually implemented as the average of three color channels (e.g., [14][117]) and processed by center-surround processes inspired by neural responses in lateral geniculate nucleus (LGN) [10] and V1 cortex. Color is implemented as red-green and blue-yellow channels inspired by color-opponent neurons in V1 cortex, or alternatively by using other color spaces such as HSV [50] or Lab [160]. Orientation is often implemented as a convolution with oriented Gabor filters or by the application of oriented masks. Motion was first used in [119] and was implemented by applying directional masks to the image (in the primate brain motion is derived by the neurons at MT and MST regions which are selective to direction of motion). Some studies have also added specific features for

directing attention like skin hue [120], face [167], horizontal line [93], wavelet [133], gist [92][93], center-bias [123], curvature [124], spatial resolution [125], optical flow [15][126], flicker [119], multiple superimposed orientations (crosses or corners) [127], entropy [129], ellipses [128], symmetry [136], texture contrast [131], above average saliency [131], depth [130], and local center-surround contrast [189]. While most models have used the features proposed by FIT [81], some approaches have incorporated other features like Difference of Gaussians (DOG) [144][141] and features derived from natural scenes by ICA and PCA algorithms [92][142]. For target search, some have employed the structural description of objects such as the histogram of local orientations [87][199]. For detailed information regarding important features in visual search and direction of attention, please refer to [118][81][82]. Factor $\mathtt{f}_{10}$, categorizes models based on features they use.

### 2.6 Stimuli and Task Type

Visual stimulus can be first distinguished as being either static (e.g., search arrays, still photographs; factor $\mathtt{f}_4$) or dynamic (e.g., videos, games; factor $\mathtt{f}_5$). Video games are interactive and highly dynamic since they do not generate the same stimuli each run and have nearly natural renderings, though they still lag behind the statistics of natural scenes and do not have the same noise distribution. The setups here are more complex, more controversial, and more computationally intensive. They also engage a large number of cognitive behaviors.

The second distinction is between synthetic stimuli (Gabor patches, search arrays, cartoons, virtual environments, games; factor $\mathtt{f}_6$) and natural stimuli (or approximations thereof, including photographs and videos of natural scenes; factor $\mathtt{f}_7$). Since humans live in a dynamic world, video and interactive environments provide a more faithful representation of the task facing the visual system than static images. Another interesting domain for studying attentional behavior, agents in virtual reality setups, can be seen in the work of Sprague and Ballard [109], who employed a realistic human agent in VR and used reinforcement learning (RL) to coordinate action selection and visual perception in a sidewalk navigation task involving avoiding obstacles, staying on the sidewalk, and collecting litter.

Factor $\mathtt{f}_8$ distinguishes among task types. The three most widely explored tasks to date in the context of attention modeling are: (1) Free viewing tasks, in which subjects are supposed to freely watch the stimuli (there is no task or question here, but many internal cognitive tasks are usually engaged), (2) Visual search tasks where subjects are asked to find an odd item or a specific object in a natural scene, and (3) Interactive tasks. In many real-world situations, tasks such as driving and playing soccer engage subjects tremendously. These complex tasks involve many subtasks such as visual search, object tracking, and focused and divided attention.

### 2.7 Evaluation Measures

So we have a model that outputs a saliency map $S$, and we have to quantitatively evaluate it by comparing it with eye movement data (or click positions) $G$. How do you compare these? We can think of them as probability distributions, and use Kullback-Leibler (KL) or Percentile metrics to measure distance between distributions. Or we can consider $S$ as a binary classifier and use signal detection theory analysis (Area Under the ROC Curve (AUC) metric) to asses the performance of this classifier. We can also think of $S$ and $G$ as random variables and use Correlation Coefficient (CC) or Normalized Scanpath Saliency (NSS) to measure their statistical relationship. Another way is to think of $G$ as a sequence of eye fixations (scanpath) and compare this sequence with the sequence of fixations chosen by a saliency model (string-edit distance).

While in principle any model might be evaluated using any measure, in Fig. 7 we list in factor $\mathtt{f}_{12}$ the measures which were used by the authors of each model. In the rest, when we use Estimated Saliency Maps (ESM $S$), we mean a saliency map of a model, and by Ground-truth Saliency Map (GSM $G$), we mean a map that is built by combining recorded eye fixations from all subjects or combining tagged salient regions by human subjects for each image.

From another perspective, evaluation measures for attention modeling can be classified into three categories: 1) point-based, 2) region-based, and 3) subjective evaluation. In point-based measures, salient points from ESMs are compared to GSMs made by combining eye fixations. Region-based measures are useful for evaluating attention models over regional saliency datasets by comparing the ESMs and labeled salient regions (GSM annotated by human subjects) [133]. In [103], subjective scores on estimated saliency maps were reported on three levels: "Good", "Acceptable", and "Failed". The problem with such subjective evaluation is that it is difficult to extend it to large-scale datasets.

In the following, we focus on explaining those metrics with more consensus from the literature and provide pointers for others (Percentile [134] and Fixation Saliency Method (FS) [131][182]) for reference.

**Kullback-Leibler (KL) Divergence**. The KL divergence is usually used to measure distance between two probability distributions. In the context of saliency, it is used to measure the distance between distributions of saliency values at human vs. random eye positions [145][77]. Let $t_i = 1 \cdots N$ be $N$ human saccades in the experimental session. For a saliency model, ESM is sampled (or averaged in a small vicinity) at the human saccade $x_{i,human}$ and at a random point $x_{i,random}$. The saliency magnitude at the sampled locations is then normalized to the range [0,1]. The histogram of these values in $q$ bins covering the range [0,1] across all saccades is then calculated. $H_k$ and $R_k$ are the fraction of points in bin $k$ for salient and random points. Finally the difference between these histograms with the (symmetric) KL divergence (A.k.a relative entropy) is:

$$KL = \frac{1}{2}\sum_{k=1}^{q}\left(H_k log\frac{H_k}{R_k} + R_k log\frac{R_k}{H_k}\right) \quad (1)$$

Models that can better predict human fixations exhibit higher KL divergence, since observers typically gaze towards a minority of regions with the highest model responses while avoiding the majority of regions with low model responses. Advantages of KL divergence over other scoring schemes [212][131] are: 1) Other measures essentially calculate the rightward shift of $H_k$ histogram relative to the $R_k$ histogram, whereas KL is sensitive to any difference between the histograms, and 2) KL is invariant to reparameterizations, such that applying any continuous monotonic nonlinearity (e.g., $S^3, \sqrt{S}, e^S$) to ESM values

$S$ does not affect scoring. One disadvantage of the KL divergence is that it does not have a well-defined upper bound — as the two histograms become completely non-overlapping, the KL divergence approaches infinity.

**Normalized Scanpath Saliency (NSS)**. The normalized scanpath saliency [134][131] is defined as the response value at the human eye position, $(x_h, y_h)$, in a model's ESM that has been normalized to have zero mean and unit standard deviation $NSS = \frac{1}{\sigma_s}(S(x_h, y_h) - \mu_S)$. Similar to the percentile measure, NSS is computed once for each saccade, and subsequently the mean and standard error are computed across the set of NSS scores. $NSS = 1$ indicates that the subjects' eye positions fall in a region whose predicted density is one standard deviation above average. Meanwhile $NSS \leq 0$ indicates that the model performs no better than picking a random position on the map. Unlike KL and percentile, NSS is not invariant to reparameterizations. Please see [134] for an illustration of NSS calculation.

**Area Under Curve (AUC)**. AUC is the area under Receiver Operating Characteristic (ROC) [195] curve. As the most popular measure in the community, ROC is used for the evaluation of a binary classifier system with a variable threshold (usually used to classify between two methods like saliency vs. random). Using this measure, the model's ESM is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated [144][167]. Human fixations are then used as ground truth. By varying the threshold, the ROC curve is drawn as the *false positive rate* vs. *true positive rate*, and the area under this curve indicates how well the saliency map predicts actual human eye fixations. Perfect prediction corresponds to a score of 1. This measure has the desired characteristic of transformation invariance, in that area under the ROC curve does not change when applying any monotonically increasing function to the saliency measure. Please see [192] for an illustration of ROC calculation.

**Linear Correlation Coefficient (CC)**. This measure is widely used to compare the relationship between two images for applications such as image registration, object recognition, and disparity measurement [196][197]. The linear correlation coefficient measures the strength of a linear relationship between two variables:

$$CC(G,S) = \frac{\sum_{x,y}(G(x,y) - \mu_G).(S(x,y) - \mu_S)}{\sqrt{\sigma_G^2.\sigma_S^2}} \qquad (2)$$

where $G$ and $S$ represent the GSM (fixation map, a map with 1's at fixation locations, usually convolved with a Gaussian) and the ESM, respectively. $\mu$ and $\sigma^2$ are the mean and the variance of the values in these maps. An interesting advantage of CC is the capacity to compare two variables by providing a single scalar value between -1 and +1. When the correlation is close to $+1/-1$ there is almost a perfectly linear relationship between the two variables.

**String Editing Distance**. To compare the regions of interest selected by a saliency model (mROI) to human regions of interest (hROI) using this measure, saliency maps and human eye movements are first clustered to some regions. Then ROIs are ordered by the value assigned by the saliency algorithm or temporal ordering of human fixations in the scanpath. The results are strings of ordered points such as: $string_h$ = "$abcfeffgdc$" and $string_s$ = "$afbffdcdf$". The string editing similarity index $S_s$ is then defined by an optimization algorithm with unit cost assigned to the three different operations: *deletion*, *insertion*, and *substitution*. Finally the sequential similarity between the two strings is defined as: $similarity = 1 - \frac{S_s}{|string_s|}$. For our example strings, above similarity is $1 - 6/9 = 0.34$ (see [198][127] for more information on string editing distance). Please see [127] for an illustration of this score.

### 2.8 Datasets

There are several eye movement datasets of still images (for studying static attention) and videos (for studying dynamic attention). In Fig. 7 we list as factor $\mathtt{f}_{13}$ some available datasets. Here we only mention those datasets that are mainly used for evaluation and comparison of attention models, though there are many other works that have gathered special-purpose data (e.g., for driving, sandwich making, and block copying [135]).

Figs. 4 and 5 show summaries of image and video eye movements datasets (For a few, labeled salient regions are available). Researchers have also used mouse tracking to estimate attention. Although this type of data is noisier, some early results show a reasonably good ground-truth approximation. For instance, Scheier and Egner [61] showed that mouse movement patterns are close to eye-tracking patterns. A web-based mouse tracking application was set up at the TCTS laboratory [110]. Other potentially useful datasets (which are not eye-movement datasets) are tagged-object datasets like PASCAL and Video LabelMe. Some attentional works have used this type of data (e.g., [116]).

## 3 Attention Models

In this section, models are explained based on their mechanism to obtain saliency. Some models fall into more than one category. In the rest of this review, we focus only on those models which have been implemented in software and can process arbitrary digital images and return corresponding saliency maps. Models are introduced in chronological order. Note that here we are more interested in models of saliency instead of those approaches that detect and segment the most salient region or object in a scene. While these models use a saliency operator at the initial stage, their main goal is not to explain attentional behavior. However, some methods have further inspired subsequent saliency models. Here, we reserve the term "saliency detection" to refer to such approaches.

### 3.1 Cognitive Models (C)

Almost all attentional models are directly or indirectly inspired by cognitive concepts. The ones that have more bindings to psychological or neurophysiological findings are described in this section.

*Itti* et al.'s basic model [14] uses three feature channels color, intensity, and orientation. This model has been the basis of later models and the standard benchmark for comparison. It has been shown to correlate with human eye movements in free-viewing tasks [131][184]. An input image is subsampled into a Gaussian pyramid and each pyramid level $\sigma$ is decomposed into channels for Red ($R$), Green ($G$), Blue ($B$), Yellow ($Y$), Intensity ($I$), and local orientations

| Study | Subjects | Dataset Size | Resolution | Viewing distance (cm) | Presentation time (s) | Description |
|---|---|---|---|---|---|---|
| Kienzle et al. [165] | 14 | 200 | 1024 x 768 | 60 | 3 | 8-bit grayscale stimuli presented on a 19-inch Iiyama CRT at full screen size corresponding to 37° × 27° of visual angle. |
| Einhauser et al. [84] | 7 | 54 | 640 x 480 | 50 | - | Overall 32,225 fixations with average fixation duration as 370±293 ms and 11.9 fixations per image. The average distance of subsequent fixation points on the screen is 127 pixels(19°). Authors restricted their analysis to 76° ×55° regions which accounts for 92% (29,725) of all fixations. Stimuli was presented using NEC LT 157 projector at resolution 1024 × 768 at 60Hz on average spanned 133 x 100cm, corresponding to 37° × 27° of visual angle. |
| Ouerhani et al. [210] | 6 | - | 640 x 480 | 70 | 5 | Age range (24-34), with normal or corrected-to-normal acuity as well as normal color vision. Stimulus presented on a 19" monitor subtending 29° × 22°. Task was "just look at the image". Eyetracker: EyeLink, SenseoMotoric Instruments GmbH. Recording at 250Hz, accuracy 0.5° – 1° accuracy with a 3×3 point grid calibration sequence. |
| Bruce and Tsotsos [144] | 20 | 120 | 681 x 511 | 75 | 4 | Images (indoor and outdoor) were presented at random with 2 s gray mask in between on a 21-inch CRT monitor. The eye tracking apparatus consisted of an ERICA workstation including Hitachi CCD camera with an IR emitting LED. Stimuli were color images and task was free viewing. **Link:** www-sop.inria.fr/members/Neil.Bruce |
| Stark and Choi [211] | 7 | 15 | - | 40 | 4 | Bright Purkinje reflection captured by a video camera. Stimulus size was 15 × 20cm yielding to 21° × 29° with the 0.5-1 degree accuracy. Images were terrain photographs, landscapes and paintings. Task was free viewing. |
| Chikkerur et al. [154] | 8 | 220 | 640 x 480 | 70 | 5 | Scenes contained cars (4.6 ± 3.8) and pedestrians (2.1 ± 2.2)- visual angle: 16 × 12. Subjects were asked to count the number of cars or pedestrians. Using an ETL 400 ISCAN, table-mounted video-based eye tracker at 240 HZ and accuracy of 0.5°. (age:18-35). Images were 100 from x and 120 from LabelMe. **Link:** http://www.sharat.org/ |
| Torralba et al. [92] | 24 | 36 | 15.8 x 11.9 | - | - | In people search task, 14 stimuli out of 36 contained no people and 22 included 1-6 people. The same set (36 indoor) images was used for painting search (17 images without any paintings and rest with 1-6 paintings) and for mug search (half without and half with 1-6 mugs). Eyetracking was performed by a Generation 5.5 SRI Dual Purkinje Image Eyetracker, sampling at 1000Hz. Color photos displayed on a NEC Multisync P750 monitor (143Hz refresh). Mean target size was 1.05%(1.24%) of the image size for people, 7.3%(7.6%) for painting and 0.5% (0.4%) for mugs. **Link:** http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/ |
| Judd et al. [166] | 15 | 1003 | Various | 48 | 3 | Images were collected from Flicker creative commons and LabelMe datasets. The longest dimension was 1024 with other ranging from 405 to 1024. There were 779 landscape images and 228 portrait images. Images were freely viewed with 1 sec gray screen between each two. Camera was recalibrated after every 50 images. First fixation was discarded. Age range: 18-35. **Link:** http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html |
| Cerf et al. [167] | 7 | 250 | 1024 x 768 | 80 | - | Eye position of subjects were acquired at 1000Hz using an Eyelink 1000 (SR Research, Osgoode, Canada). The task had three phases: 1) free viewing, 2) searching for face, an object, banana, cell phone, toy car, etc shown by a probe image, and 3) 100 image recognition memory task where subjects had to answer with y/n whether they had seen the image before. Stimuli subtended 28° × 21° of visual angle. **Link:** http://www.fifadb.com/ |
| Peters et al. [134] | 12 | 100/ class | - | 75 | - | ISCAN Inc eye tracker was used to sample eye movements at 120Hz. Age range: 18-25; four did free-viewing over (outdoor photos, overhead satellite imagery, and fractals). Another 4 did free-viewing over involving Gabor snakes and Gabor arrays. Seven subjects did a contour detection task. Resolution was 1000 × 1000 to 1536 × 1024 subtending a visual angle of 15.8° × 15.8° to 16.2° × 25°. **Link:** http://ilab.usc.edu |
| Reinagel and Zador [212] | 5 | 77 | 640 x 480 | 79 | 10 | Images were 69 nature scenes, 38 man-made objects such as buildings, 17 animals or humans and 8 synthetics. An RK-416 infrared Pupil Tracking System and a 21-inch monitor was used. The whole image subtended 28° × 21° of visual angle. Subjects were instructed to "Study the images". Estimated tracking error was 0.5°. **Link:** http://zadorlab.cshl.edu/ |
| Hwang and Pomplun [87] | 30 | 160 | 1280 x 1024 | - | 10 | Age range: 19-40. Stimuli were 160 photographs (1280 × 1024) real-world scenes including landscapes, home interiors, and city scenes and covered 20° × 20° of visual angle. An SR research EyeLink II system. Stimuli presented on 19-inch Del P992 monitor (85Hz refresh rate), the whole image subtended 28° × 21°. **Link:** http://www.cs.umb.edu/~marc/ |
| Kootstra et al., [136] | 31 | 99 | 1024 x 768 | 70 | - | Eyelink head-mounted eye tracking (SR research) was used and was recalibrated before each session. Age range: 17-32. Task was free viewing. Stimuli were: 12 Animals, 12 Automan, 16 Buildings, 20 flowers, 41 natural scenes and were shown on a 18-inch CRT monitor (36 × 27 cm). **Link:** http://www.csc.kth.se/~kootstra/ |
| Tatler [123] | 14 | 48 | 800 x 600 | 60 | - | Eyelink-I eye tracker was used. Subjetcs had normal or corrected to normal vision with age range 17-32. Image subtended 30° × 22° and were presented on a 17-inch SVGA color monitor (74 Hz refresh). Task was free viewing. **Link:** http://www.activevisionlab.org/ |
| Engmann et al., [182] | 8 | 90 | 1280 x 1024 | 85 | - | Subjetcs had normal or corrected-to-normal vision and normal color vision with age range 20-27 (avg: 22.3). Stimuli were presented on a 19.7" EIZO FlexScan F77S CRT monitor (100 Hz refresh). Natural scenes selected from the Zurich natural image database (Einhauser et al. [99]) which only rarely contain isolated nameable objects or man-made artifacts at resolution 2048 × 1536. Image subtended 26° × 18° 17-inch SVGA color monitor. Task was free viewing. Eye tracker was Eyelink-2000 (SR Research Ltd. Canada) with 13 point calibration. |
| Engelke et al. [213] | 30 | 7 | 512 x 512 | 60 | 8 | Images were 4 human faces ("Barbara"), 1 "Glodhill" face (gurilla) and 1 "Peppers" images. Eye tracker was EyeTech TM3 and task was free viewing. Each image was presented for 8 sec with a gray screen with central fixation in between. |
| Le Meur et al. [41] | 40 | 46 | 800 x 600 | * | - | Stimuli were 46 degraded versions of 10 color images using spatial filtering. Task was free viewing. Eye tracker was made by Cambridge Research Corporation. Viewing distance was four times the TV monitor height. **Link:** http://www.irisa.fr/temics/staff/lemeur |
| Ehinger et al. [87] | 14 | 912 | 800 x 600 | 75 | 15 | Stimuli were color images (half with a pedestrian) with resolution 800 × 600 and were shown on a 21-inch CRT monitor with resolution 1024 × 768 and refresh rate 100Hz. A 240 Hz ISCAN RK-464 video-based eye tracker was used for recording. The task was to decide wether a pedestrain is in the scene or not. **Link:** http://cvcl.mit.edu/searchmodels/ |
| Rajashekar et al. [174] | 29 | 101 | 1042 x 768 | 134 | - | Subjects were 18 males, 11 females with mean age of 27. Eye tracker was made by Image Systems Corp, MN. Grayscale images were shown on a 21-inch grayscale gamma corrected monitor with resolution 1024 × 768. The task was free viewing. **Link:** http://live.ece.utexas.edu/research/doves/ |

Fig. 4. Some benchmark eye movement datasets over still images often used to evaluate visual attention models.

($O_\theta$). From these channels, center-surround "feature maps" $f_l$ for different features $l$ are constructed and normalized. In each channel, maps are summed across scale and normalized again:

$$\begin{aligned} f_l &= \mathcal{N}\left(\sum_{c=2}^{4}\sum_{s=c+3}^{c+4} f_{l,c,s}\right), \forall l \in L_I \cup L_C \cup L_O \\ L_I &= \{I\}, L_C = \{RG, BY\}, L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \end{aligned} \quad (3)$$

These maps are linearly summed and normalized once more to yield the "conspicuity maps":

$$C_I = f_I, C_C = \mathcal{N}(\sum_{l\in L_C} f_l), C_O = \mathcal{N}(\sum_{l\in L_O} f_l) \quad (4)$$

Finally, conspicuity maps are linearly combined once more to generate the saliency map: $S = \frac{1}{3}\sum_{k\in\{I,C,O\}} C_k$.

There are at least four implementations of this model: iNVT by Itti [14], Saliency Toolbox (STB) by Walther [35], VOCUS by Frintrop [50], and a Matlab code by Harel [121]. In [119], this model was extended by adding motion and flicker contrasts to video domain. Zhaoping Li [170], introduced a neural implementation for saliency map in V1 area that can also account for search difficulty in pop-out and conjunction search tasks.

*Le Meur* et al. [41] proposed an approach for bottom-up saliency based on the structure of the human visual system (HVS). Contrast sensitivity functions, perceptual decompo-

| Dataset | Features | Feature Value |
|---|---|---|
| CRCNS - ORIG [145] | C | 50 clips (0:06-1:30 min each), ~25 min total, ~6GB for 46K frames |
| | S | 8 (3 female, 5 male) subjects with normal corrected vision, Ages 23-32, From mixed ethnicities |
| | T | "Follow main actors and actions, try to understand overall what happens in each clip." |
| | ST | Complex video stimuli involving TV programs, outdoor sceces, video games Outdoor day & night, parks, crowds, rooftop bar. etc. |
| | D | ISCAN RK-464 eye tracker, 240 HZ recording, 9 point calibration after every 5 clips, 640× 480 resolution at 60.27HZ doublescan, 33.185ms/movie frame, (x,y) of each saccade |
| | L | http://crcns.org/data-sets/eye/eye-1 |
| CRCNS - MTV [145] | C | 50 video clips (4-7 subjects on each video clip) |
| | S | 8 subjects different from subjects of CRCNS |
| | D | This dataset was created by cutting video clips of CRCNS into 1-3s "clipets" and reassembling those clippets in random order. Other aspects were the same as the original dataset. |
| | L | http://crcns.org/data-sets/eye/eye-1 |
| Jia Li et al. [133] | C | 431 videos with total length of 7.5 hours, 764,806 frames in total with 62,356 key frames |
| | S | 23 (17 male and 4 female) subjects with age range between 21-37 |
| | ST | 6 genres: documentary, ad, cartoon, news, movie and surveillance |
| | D | 10-23 subjects per each clip were assigned to manually label the salient regions with one or multiple rectangles from key frames. Drawback with this dataset is rectangular labeling but this may be resolved with segmentation, inefficiney to evaluate whatever |
| | L | http://www.jdl.ac.cn/user/jiali/ |
| Peters and Itti [101] | C | 24 game-play sessions, ~185 GB for 216K frames, 8,449 saccades of ampitude 2o or more |
| | S | 5(3 male, 2 female) subjetcs with normal corrected vision |
| | T | "Play 4 or 5 five-minute segments of the Nintendo GameCube games" |
| | ST | Games include Mario Kart, Wave Race, Super Mario Sunshine, Hulk and Pac Man World. |
| | D | Subjects were seated viewing distance of 80 cm (28° × 21° usable field of view) Stimuli were presented on a 22" computer monitor (LaCie Corp; 640 × 480, 75 HZ refresh, mean screen luminance 30cd/m2, room luminance 4 cd/m2) ISCAN RK-464 eye tracker, 240 HZ recording, 9 point calibration after before game segment Frames were grabbed using a dual-CPU Linux computer with SCHED FIFO schedulingto ensure micorsecond accurate timing. |
| | L | http://ilab.usc.edu/rjpeters/ |
| Shic and Scassellati [74] | C | 2 clips, 10, young adults, normal and mildly mentally retarded |
| | T | "One minute long clips from back and white movie "Who's afraid of Virgnia Woolf" |
| | D | A head mounted eye-tracker (ISCAN Inc.) was used. The eye tracker employs dark pupil- corneal reflection video-occulography and had accuracy within ±0.3o over a horizontal and range of ±20o, with a sampling rate of 60 Hz. The subjects sat 63.5 cm from the 48.3 cm screen on which the movie was shown at a resolution of 640 × 480 pixels. |
| | L | http://sites.google.com/site/fredshic/home |
| Marat et al. [49] | C | 53 short video clips (25 fps, 720 × 576 pixels ), 1700 frames |
| | S | 15 (3f,12m) subjects with age range 23-40 and had normal or corrected to normal vision |
| | ST | Each clip ~ 1-3sec long, 324 clip snippets. There was not a particular task or question. TV shows, TV news, animated movies, commercials, sport and music. Indoor, out-door, day-time, night-time) |
| | D | The clip snippets were strung to form 20 clips of 30 seconds (30.20 ± 0.61). Eye positions were recorded at 500 Hz (20 eye positions per frame for two eyes) using a Eyelink II (SR Research). Participants were positioned with their chin supported on a 21" color monitor (75 HZ) at a viewing distance of 57cm (40° × 30° usable field of view). A calibration was carried out at every five stimuli and a control drfit was done before each stimuli. |
| | L | http://start1g.ovh.net/~qgsmabaq/sophie/index.php |
| Le Meur et al. [138] | C | 7 clips (25 Hz, 352 × 288 pixels ), 2451 frames, Each clip ~ 4.5-33.8 sec long |
| | S | 17-27 subject for different clips with normal or corrected to normal vision |
| | T | Free viewing |
| | ST | Faces, sporting events, audiencesm, landscape, logos, incrustations, low and high satiotemporal |
| | D | Dual-Purkinje eye tracker from Cambridge Research Corporation. Sampling frequency was 50Hz. CRT display 800 × 600 pixels, 25° × 27°. Distance to screen was 81 cm. |
| | L | http://www.irisa.fr/temics/staff/lemeur |

**C**: Clips; **S**: Subjects; **T**: Task; **ST**: Stimuli Type; **D:** Description; **L:** Link

Fig. 5. Some benchmark eye movement datasets over video stimuli for evaluating visual attention prediction.

sition, visual masking, and center-surround interactions are some of the features implemented in this model. Later, Le Meur *et al.* [138] extended this model to spatio-temporal domain by fusing achromatic, chromatic and temporal information. In this new model, early visual features are extracted from the visual input into several separate parallel channels. A feature map is obtained for each channel, then a unique saliency map is built from the combination of those channels. The major novelty proposed here lies in the inclusion of the temporal dimension as well as the addition of a coherent normalization scheme.

*Navalpakkam and Itti* [51] modeled visual search as a top-down gain optimization problem by maximizing the signal-to-noise ratio (SNR) of the target vs. distractors instead of learning explicit fusion functions. That is, they learned linear weights for feature combination by maximizing the ratio between target saliency and distractor saliency.

*Kootstra* et al. [136] developed three symmetry-saliency operators and compared them with human eye tracking data. Their method is based on the isotropic symmetry and radial symmetry operators of Reisfeld *et al.* [137] and the color symmetry of Heidemann [64]. Kootstra *et al.* extended these operators to multi-scale symmetry-saliency models. The authors showed that their model performs significantly better on symmetric stimuli compared to the Itti *et al.* [14].

*Marat* et al. [104] proposed a bottom-up approach for spatio-temporal saliency prediction in video stimuli. This model extracts two signals from the video stream corresponding to parvocellular and magnocellular cells of the retina. From these signals, two static and dynamic saliency maps are derived and fused into a spatio-temporal map. Prediction results of this model were better for the first few frames of each clip snippet.

*Murray* et al. [200] introduced a model based on a low

level vision system in three steps: 1) visual stimuli are processed according to what is known about the early human visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition), 2) a simulation of the inhibition mechanisms present in cells of the visual cortex normalize their response to stimulus contrast, and 3) information is integrated at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs.

Cognitive models have the advantage of expanding our view of biological underpinnings of visual attention. This further helps understanding computational principles or neural mechanisms of this process as well as other complex dependent processes such as object recognition.

### 3.2 Bayesian Models (B)

Bayesian modeling is used for combining sensory evidence with prior constraints. In these models, prior knowledge (e.g., scene context or gist) and sensory information (e.g., target features) are probabilistically combined according to Bayes' rule (e.g., to detect an object of interest).

*Torralba* [92] and *Oliva* et al. [140] proposed a Bayesian framework for visual search tasks. Bottom-up saliency is derived from their formulation as $\frac{1}{p(f|f_G)}$ where $f_G$ represents a global feature that summarizes the probability density of presence of the target object in the scene, based on analysis of the scene gist. Following the same direction, Ehinger *et al.* [87] linearly integrated three components (bottom-up saliency, gist, and object features) for explaining eye movements in looking for people in a database of about 900 natural scenes.

*Itti and Baldi* [145] defined surprising stimuli as those which significantly change beliefs of an observer. This is modeled in a Bayesian framework by computing the KL divergence between posterior and prior beliefs. This notion is applied both over space (surprise arises when observing image features at one visual location affects the observer's beliefs derived from neighboring locations) and time (surprise then arises when observing image features at one point in time affects beliefs established from previous observations).

*Zhang* et al. [141] proposed a definition of saliency, known as SUN: Saliency Using Natural statistics, by considering what the visual system is trying to optimize when directing attention. The resulting model is a Bayesian framework in which bottom-up saliency emerges naturally as the self-information of visual features, and overall saliency (incorporating top-down information with bottom-up saliency) emerges as the point-wise mutual information between local image features and the search target's features when searching for a target. Since this model provides a general framework for many models, we describe it in more detail.

SUN's formula for bottom-up saliency is similar to the work of Oliva *et al.* [140], Torralba [92], and Bruce and Tsotsos [144], in that they are all based on the notion of self-information (local information). However, differences between current image statistics and natural statistics lead to radically different kinds of self-information. Briefly, the motivating factor for using self-information with the statistics of the current image is that a foreground object is likely to have features that are distinct from those of the background. Since targets are observed less frequently than background during an organism's lifetime, rare features are more likely to indicate targets.

Let $Z$ denote a pixel in the image, $C$ whether or not a point belongs to a target class and $L$ the location of a point (pixel coordinates). Also, let $F$ be the visual features of a point. Having these, the saliency $s_z$ of a point $z$ is defined as $P(C = 1|F = f_z, L = l_z)$ where $f_z$ and $l_z$ are the feature and location of $z$. Using the Bayes rule and assuming that features and locations are independent and conditionally independent given $C = 1$, then saliency of a point is:

$$log\ s_z = -log\ P(F = f_z) + log\ P(F = f_z|C = 1) + log\ P(C = 1|L = l_z) \quad (5)$$

The first term at the right side is the self-information (bottom-up saliency) and it depends only on the visual features observed at the point $Z$. The second term on the right is the log-likelihood which favors feature values that are consistent with prior knowledge of the target (e.g., if the target is known to be green the log-likelihood will take larger values for a green point than for a blue point). The third term is the location prior which captures top-down knowledge of the target's location and is independent of visual features of the object. For example, this term may capture knowledge about some target being often found in the top-left quadrant of an image.

*Zhang* et al. [142] extended the SUN model to dynamic scenes by introducing temporal filters (Difference of Exponentials) and fitting a generalized Gaussian distribution to the estimated distribution for each filter response. This was implemented by first applying a bank of spatio-temporal filters to each video frame, then for any video, the model calculates its features and estimates the bottom-up saliency for each point. The filters were designed to be both efficient and similar to the human visual system. The probability distributions of these spatio-temporal features were learned from a set of videos from natural environments.

*Jia Li* et al. [133] presented a Bayesian multi-task learning framework for visual attention in video. Bottom-up saliency modeled by multi-scale wavelet decomposition was fused with different top-down components trained by a multi-task learning algorithm. The goal was to learn task-related "stimulus-to-saliency" functions, similar to [101]. This model also learns different strategies for fusing bottom-up and top-down maps to obtain the final attention map.

*Boccignone* [55] addressed joint segmentation and saliency computation in dynamic scenes, using a mixture of Dirichlet processes as a basis for object-based visual attention. He also proposed an approach for partitioning a video into shots based on a foveated representation of a video.

A key benefit of Bayesian models is their ability to learn from data and their ability to unify many factors in a principled manor. Bayesian models can, for example, take advantage of the statistics of natural scenes or other features that attract attention.

### 3.3 Decision Theoretic Models (D)

The decision-theoretic interpretation states that perceptual systems evolve to produce decisions about the states of the surrounding environment that are optimal in a decision theoretic sense (e.g., minimum probability of error). The overarching point is that visual attention should be driven by optimality with respect to the end task.

*Gao and Vasconcelos* [146] argued that for recognition, salient features are those that best distinguish a class of interest from all other visual classes. They then defined top-down attention as classification with minimal expected error. Specifically, given some set of features $F = \{F_1, \cdots, F_d\}$, a location $l$ and a class label $C$ with $C_l = 0$ corresponding to samples drawn from the surround region and $C_l = 1$ corresponding to samples drawn from a smaller central region centered at $l$, the judgment of saliency then corresponds to a measure of mutual information, computed as $I(F, C) = \sum_{i=1}^{d} I(F_i, C)$. They used DOG and Gabor filters, measuring the saliency of a point as the KL divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region. In [185], the same authors used this framework for bottom-up saliency by combining it with center-surround image processing. They also incorporated motion features (optical flow) between pairs of consecutive images to their model to account for dynamic stimuli. They adopted a dynamic texture model using a Kalman filter in order to capture the motion patterns in dynamic scenes.

Here we show the Bayesian computation of (5) is a special case of the Decision theoretic model. Saliency computation in the entire decision theoretic approach boils down to calculating the target posterior probability $P(C = 1|F = f_z)$ (the output of their simple cells [215]). By applying Bayesian rule, we have:

$$P(C_l = 1|F_l = f_z) = \sigma\Big(log\frac{P(F_l = f_z|C_l = 1)P(C_l = 1)}{P(F_l = f_z|C_l = 0)P(C_l = 0)}\Big) \quad (6)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The log likelihood ratio inside the sigmoid can be trivially written (using the independence assumptions of [141]) as:

$$-logP(F = f_z|C = 0) + logP(F = f_z|C = 1) + \frac{P(C = 1|L = l_z)}{P(C = 0|L = l_z)} \quad (7)$$

which is the same as (5) under the following assumptions: 1) $P(F = f_z|C = 0) = P(F = f_z)$ and 2) $P(C = 0|L = l_z) = K$, for some constant $K$. Assumption 1 states that the feature distribution in the absence of the target is the same as the feature distribution for the set of natural images. Since the overwhelming majority of natural images do not have the target, this is really not much of an assumption. The two distributions are virtually identical. Assumption 2 simply states that the absence of the target is equally likely in all image locations. This, again, seems like a very mild assumption.

Because of above connections, both Decision theoretic and Bayesian approaches have a biologically plausible implementation, which has been extensively discussed by Vasconcelos and colleagues [223][147][215]. The Bayesian methods can be mapped to a network with a layer of simple cells and the decision theoretic models to a network with a layer of simple and a layer of complex cells. The simple cell layer in fact can also implement AIM [144] and Rosenholtz [191] models in Section 3.4, Elazary and Itti [90], and probably some more. So, while these models have not been directly derived from biology, they can be implemented as cognitive models.

*Gao and Vasconcelos* [147] used discriminant saliency model for visual recognition and showed good performance on PASCAL 2006 dataset.

*Mahadevan and Vasconcelos* [105] presented an unsupervised algorithm for spatio-temporal saliency based on biological mechanisms of motion-based perceptual grouping. It is an extension of the discriminant saliency model [146]. Combining center-surround saliency with the power of dynamic textures made their model applicable to highly dynamic backgrounds and moving cameras.

In *Gu* et al. [148], an activation map was first computed by extracting primary visual features and detecting meaningful objects from the scene. An adaptable retinal filter was applied to this map to generate "regions of interest" (ROIs whose locations correspond to these activation peaks and whose sizes were estimated by an iterative adjustment algorithm). The focus of attention was moved serially over the detected ROIs by a decision theoretic mechanism. The generated sequence of eye fixations was determined from a perceptual benefit function based on perceptual costs and rewards, while the time distribution of different ROIs was estimated by memory learning and decaying.

Decision theoretic models have been very successful in computer vision applications such as classification while achieving high accuracy in fixation prediction.

### 3.4 Information Theoretic Models (I)

These models are based on the premise that localized saliency computation serves to maximize information sampled from one's environment. They deal with selecting the most informative parts of a scene and discarding the rest.

*Rosenholtz* [191][193] designed a model of visual search which could also be used for saliency prediction over an image in free-viewing. First, features of each point, $p_i$, are derived in an appropriate uniform feature space (e.g., uniform color space). Then, from the distribution of the features, mean, $\mu$, and covariance, $\sum$, of distractor features are computed. The model then defines target saliency as the Mahalanobis distance, $\Delta$, between the target feature vector, $T$, and the mean of the distractor distribution, where $\Delta^2 = (T - \mu)^{'} \sum^{-1} (T - \mu)$. This model is similar to [92][141][160] in the sense that it estimates $1/P(x)$ (rarity of a feature or self-information) for each image location $x$. This model also underlies a clutter measure of natural scenes (same authors [189]). An online version of this model is available at [194].

*Bruce and Tsotsos* [144] proposed the AIM model (Attention based on Information Maximization) which uses Shannon's self-information measure for calculating saliency of image regions. Saliency of a local image region is the information that region conveys relative to its surroundings. Information of a visual feature $X$ is $I(X) = -log\ p(X)$, which is inversely proportional to the likelihood of observing $X$ (i.e., $p(X)$). To estimate $I(X)$, the probability density function $p(X)$ must be estimated. Over RGB images, considering a local patch of size $M \times N$, $X$ has the high dimensionality of $3 \times M \times N$. To make the estimation of $p(X)$ feasible, they used ICA to reduce the dimensionality of the problem to estimating $3 \times M \times N$ 1D probability density functions. To find the bases of ICA, they used a large sample of RGB patches drawn from natural scenes. For a given image, the 1D pdf for each ICA basis vector is first computed using non-parametric density estimation. Then, at each image location, the probability of observing the RGB values in a local image patch is the product of the corresponding ICA basis likelihoods for that patch.

*Hou and Zhang* [151] introduced the Incremental Coding Length (ICL) approach to measure the respective entropy gain of each feature. The goal was to maximize the entropy of the sample visual features. By selecting features with large coding length increments, the computational system can achieve attention selectivity in both dynamic and static scenes. They proposed ICL as a principle by which energy is distributed in the attention system. In this principle, the salient visual cues correspond to unexpected features. According to the definition of ICL, these features may elicit entropy gain in the perception state and are therefore assigned high energy.

*Mancas* [152] hypothesized that attention is attracted by minority features in an image. The basic operation is to count similar image areas by analyzing histograms which makes this approach closely related to Shannon's self-information measure. Instead of comparing only isolated pixels it takes into account the spatial relationships of areas surrounding each pixel (e.g., mean and variance). Two types of rarity models are introduced: Global and Local. While global rarity considers uniqueness of features over entire image, some image details may still appear salient due to local contrast or rarity. Similar to the center-surround ideas of [14], they used a multi-scale approach for the computation of local contrast.

*Seo and Milanfar* [108] proposed the Saliency prediction by Self-Resemblance (SDSR) approach. First a local image structure at each pixel is represented by a matrix of local descriptors (local regression kernels), which are robust in the presence of noise and image distortions. Then, matrix cosine similarity (a generalization of cosine similarity) is employed to measure the resemblance of each pixel to its surroundings. For each pixel, the resulting saliency map represents the statistical likelihood of its feature matrix $F_i$ given the feature matrices $F_j$ of the surrounding pixels:

$$s_i = \frac{1}{\sum_{j=1}^{N} exp(\frac{-1+\rho(F_i,F_j)}{\sigma^2})} \tag{8}$$

where $\rho(F_i, F_j)$ is the matrix cosine similarity between two feature maps $F_i$ and $F_j$, and $\sigma$ is a local weighting parameter. The columns of local feature matrices represent the output of local steering kernels which are modeled as:

$$K(x_l - x_i) = \frac{\sqrt{det(C_i)}}{h^2} exp\left\{\frac{(x_l - x_i)^T C_l (x_l - x_i)}{-2h^2}\right\} \tag{9}$$

where $l = 1, ..., P$, $P$ is the number of the pixels in a local window, $h$ is a global smoothing parameter, and the matrix $C_l$ is a covariance matrix estimated from a collection of spatial gradient vectors within the local analysis window around a sampling position $x_l = [x_1, x_2]_l^T$.

*Yin Li* et al. [171] proposed a visual saliency model based on conditional entropy for both image and video. Saliency was defined as the minimum uncertainty of a local region given its surrounding area (namely the minimum conditional entropy), when perceptional distortion is considered. They approximated the conditional entropy by the lossy coding length of multivariate Gaussian data. The final saliency map was accumulated by pixels and further segmented to detect the proto-objects. Yan *et al.* [186] proposed a newer version of this model by adding a multi resolution scheme to it.

*Wang* et al. [201], introduced a model to simulate human saccadic scanpaths on natural images by integrating three related factors guiding eye movements sequentially: 1) reference sensory responses, 2) fovea-periphery resolution discrepancy, and 3) visual working memory. They compute three multi-band filter response maps for each eye movement which are then combined into multi-band residual filter response maps. Finally, they compute residual perceptual information (RPI) at each location. The next fixation is selected as the location with the maximal RPI value.

### 3.5 Graphical Models (G)

A graphical model is a probabilistic framework in which a graph denotes the conditional independence structure between random variables. Attention models in this category treat eye movements as a time series. Since there are hidden variables influencing the generation of eye movements, approaches like Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN), and Conditional Random Fields (CRF) have been incorporated.

*Salah* et al. [52] proposed an approach for attention and applied it to handwritten digit and face recognition. In the first step (Attentive level), a bottom-up saliency map is constructed using simple features. In the intermediate level "what" and "where" information is extracted by dividing the image space into uniform regions and training a single-layer perceptron over each region in a supervised manner. Eventually this information is combined at the associative level with a discrete Observable Markov Model (OMM). Regions visited by a fovea are treated as states of the OMM. An inhibition of return allows the fovea to focus on the other positions in the image.

*Liu* et al. [43] proposed a set of novel features and adopted a Conditional Random Field to combine these features for salient object detection on their regional saliency dataset. Later, they extended this approach to detect salient object sequences in videos [48]. They presented a supervised approach for salient object detection, formulated as an image segmentation problem using a set of local, regional and global salient object features. A CRF was trained and evaluated on a large image database containing 20,000 labeled images by multiple users.

*Harel* et al. [121] introduced Graph-Based Visual Saliency (GBVS). They extract feature maps at multiple spatial scales. A scale-space pyramid is first derived from image features: intensity, color, and orientation (similar to Itti *et al.* [14]). Then, a fully-connected graph over all grid locations of each feature map is built. Weights between two nodes are assigned proportional to the similarity of feature values and their spatial distance. The dissimilarity between two positions $(i, j)$ and $(p, q)$ in the feature map, with respective feature values $M(i, j)$ and $M(p, q)$, is defined as:

$$d((i,j) \parallel (p,q)) = |log\frac{M(i,j)}{M(p,q)}| \tag{10}$$

The directed edge from node $(i, j)$ to node $(p, q)$ is then assigned a weight proportional to their dissimilarity and their distance on lattice $M$:

$$w((i,j),(p,q)) = d((i,j) \parallel (p,q)).F(i-p, j-q)$$
$$\text{where } F(a,b) = exp\Big(-\frac{a^2+b^2}{2\sigma^2}\Big) \tag{11}$$

The resulting graphs are treated as Markov chains by normalizing the weights of the outbound edges of each node

to 1 and by defining an equivalence relation between nodes and states, as well as between edge weights and transition probabilities. Their equilibrium distribution is adopted as the activation and saliency maps. In the equilibrium distribution, nodes that are highly dissimilar to surrounding nodes will be assigned large values. The activation maps are finally normalized to emphasize conspicuous detail, and then combined into a single overall map.

*Avraham* et al. [153] introduced the E-saliency (Extended saliency) model by utilizing a graphical model approximation to extend their static saliency model based on self similarities. The algorithm is essentially a method for estimating the probability that a candidate is a target. The E-Saliency algorithm is as follows: 1) Candidates are selected using some segmentation process, 2) The preference for a small number of expected targets (and possibly other preferences) is used to set the initial (prior) probability for each candidate to be a target, 3) The visual similarity is measured between every two candidates to infer the correlations between the corresponding labels, 4) Label dependencies are represented using a Bayesian network, 5) The $N$ most likely joint label assignments are found, and 6) Saliency of each candidate is deduced by marginalization.

*Pang* et al. [102] presented a stochastic model of visual attention based on the signal detection theory account of visual search and attention [155]. Human visual attention is not deterministic and people may attend to different locations on the same visual input at the same time. They proposed a dynamic Bayesian network to predict where humans typically focus in a video scene. Their model consists of four layers. In the first layer, a saliency map (Itti's) is derived that shows the average saliency response in each location in a video frame. Then in the second layer, a stochastic saliency map converts the saliency map into natural human responses through a Gaussian state space model. As to the third layer, an eye movement pattern controls the degree of overt shifts of attention through a Hidden Markov Model and finally an eye focusing density map predicts positions that people likely pay attention to based on the stochastic saliency map and eye movement patterns. They reported a significant improvement in eye fixation detection over previous efforts at the cost of decreased speed.

*Chikkerur* et al. [154] proposed a model similar to the model of Rao *et al.* [217] based on assumptions that the goal of the visual system is to know what is where and that visual processing happens sequentially. In this model, attention emerges as the inference in a Bayesian graphical model which implements interactions between ventral and dorsal areas. This model is able to explain some physiological data (neural responses in ventral stream (V4 and PIT) and dorsal stream (LIP and FEF)) as well as psychophysical data (human fixations in free viewing and search tasks).

Graphical models could be seen as a generalized version of Bayesian models. This allows them to model more complex attention mechanisms over space and time which results in good prediction power (e.g., [121]). The drawbacks lie in model complexity, especially when it comes to training and readability.

### 3.6 Spectral Analysis Models (S)

Instead of processing an image in the spatial domain, models in this category derive saliency in the frequency domain.

*Hou and Zhang* [150] developed the spectral residual saliency model based on the idea that similarities imply redundancies. They propose that statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects become conspicuous. Given an input image $I(x)$, amplitude $\mathcal{A}(f)$ and phase $\mathcal{P}(f)$ are derived. Then, the log spectrum $\mathcal{L}(f)$ is computed from the down-sampled image. From $\mathcal{L}(f)$, the spectral residual $\mathcal{R}(f)$ can be obtained by multiplying $\mathcal{L}(f)$ with $h_n(f)$ which is an $n \times n$ local average filter and subtracting the result from itself. Using the inverse Fourier transform, they construct the saliency map in the spatial domain. The value of each point in the saliency map is then squared to indicate the estimation error. Finally, they smooth the saliency map with a Gaussian filter $g(x)$ for better visual effect. The entire process is summarized below:

$$\begin{aligned} \mathcal{A}(f) &= \mathcal{R}\Big(\mathcal{F}[I(x)]\Big), \\ \mathcal{P}(f) &= \varphi\Big(\mathcal{F}[I(x)]\Big), \\ \mathcal{L}(f) &= log\Big(\mathcal{A}(f)\Big), \\ \mathcal{R}(f) &= \mathcal{L}(f) - h_n(f) * \mathcal{L}(f), \\ \mathcal{S}(x) &= g(x) * \mathcal{F}^{-1}\Big[exp\Big(\mathcal{R}(f) + \mathcal{P}(f)\Big)\Big]^2 \end{aligned} \tag{12}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and Inverse Fourier Transforms, respectively. $\mathcal{P}$ denotes the phase spectrum of the image, and is preserved during the process. Using a threshold they find salient regions called proto objects for fixation prediction. As a testament to its conceptual clarity, residual saliency could be computed in 5 lines of Matlab code [187]. But note that these lines exploit complex functions that has long implementations (e.g., $\mathcal{F}$ and $\mathcal{F}^{-1}$).

*Guo* et al. [156] showed that incorporating the phase spectrum of the Fourier transform instead of the amplitude transform leads to better saliency predictions. Later, Guo *et al.* [157] proposed a quaternion representation of an image combining intensity, color, and motion features. They called this method "phase spectrum of quaternion Fourier transform (PQFT)" for computing spatio-temporal saliency and applied it to videos. Taking advantage of the multi-resolution representation of the wavelet, they also proposed a foveation approach to improve coding efficiency in video compression.

*Achanta* et al. [158] implemented a frequency-tuned approach to salient region detection using low-level features of color and luminance. First, the input RGB image $I$ is transformed to CIE $Lab$ color space. Then, the scalar saliency map $S$ for image $I$ is computed as: $S(x,y) = \|I_\mu - I_{\omega_{hc}}\|$ where $I_\mu$ is the arithmetic mean image feature vector, $I_{\omega_{hc}}$ is a Gaussian blurred version of the original image using a $5 \times 5$ separable binomial kernel, $\|.\|$ is the $L_2$ norm (Euclidean distance), and $x, y$ are the pixel coordinates.

*Bian and Zhang* [159] proposed the Spectral Whitening (SW) model based on the idea that visual system bypasses the redundant (frequently occurring, non-informative) features while responding to rare (informative) features. They used spectral whitening as a normalization procedure in the construction of a map that only represents salient features and localized motion while effectively suppressing redundant (non-informative) background information and ego-motion. First, a grayscale input image $I(x,y)$ is low-pass fil-

tered and subsampled. Next, a windowed Fourier transform of the image is calculated as: $f(u,v) = F[w(I(x,y))]$, where $F$ denotes the Fourier transform and $w$ is a windowing function. The normalized (flattened or whitened) spectral response $\left(n(u,v) = f(u,v)/\|f(u,v)\|\right)$ is transformed into the spatial domain through the inverse Fourier transform ($F^{-1}$) squared to emphasize salient regions. Finally it is convolved with a Gaussian low-pass filter $g(u,v)$ to model the spatial pooling operation of complex cells: $S(x,y) = g(u,v) * \|F^{-1}[n(u,v)]\|^2$.

Spectral analysis models are simple to explain and implement. While still very successful, biological plausibility of these models is not very clear.

### 3.7 Pattern Classification Models (P)

Machine learning approaches have also been used in modeling visual attention by learning models from recorded eye-fixations or labeled salient regions. Typically, attention control works as a "stimuli-saliency" function to select, reweight, and integrate the input visual stimuli. Note that these models may not be purely bottom-up since they use features that guide top-down attention (e.g., faces or text).

*Kienzle* et al. [165] introduced a non-parametric bottom-up approach for learning attention directly from human eye tracking data. The model consists of a nonlinear mapping from an image patch to a real value, trained to yield positive outputs on fixations, and negative outputs on randomly selected image patches. The saliency function is determined by its maximization of prediction performance on the observed data. A support vector machine (SVM) was trained to determine the saliency using the local intensities. For videos, they proposed to learn a set of temporal filters from eye-fixations to find the interesting locations.

The advantage of this approach is that it does not need a priori assumptions about features that contribute to salience or how these features are combined to a single salience map. Also this method produces center-surround operators analogous to receptive fields of neurons in early visual areas (LGN and V1).

*Peters and Itti* [101] trained a simple regression classifier to capture the task-dependent association between a given scene (summarized by its gist) and preferred locations to gaze at while human subjects were playing video games. During testing of the model, the gist of a new scene is computed for each video frame, and is used to compute the top-down map. They showed that a point-wise multiplication of bottom-up saliency with the top-down map learned in this way results in higher prediction performance.

*Judd* et al. [166], similar to Kienzle *et al.* [165], trained a linear SVM from human fixation data using a set of low, mid, and high-level image features to define salient locations. Feature vectors from fixated locations and random locations, were assigned $+1$ and $-1$ class labels, respectively. Their results over a dataset of 1003 images observed by 15 subjects (gathered by the same authors) show that combining all aforementioned features plus distance from image center produces the best eye fixation prediction performance.

As available eye movement data increases and with wider spread of eye tracking devices supporting gathering mass data, these models are becoming popular. This however, makes models data-dependent thus influencing fair model comparison, slow, and to some extent, black-box.

### 3.8 Other Models (O)

Some other attention models that do not fit into our categorization are discussed below.

*Ramstrom and Christiansen* [168] introduced a saliency measure using multiple cues based on game theory concepts inspired by the selective tuning approach of Tsotsos *et al.* [15]. Feature maps are integrated using a scale pyramid where the nodes are subject to trading on a market and the outcome of the trading represents the saliency. They use the spot-light mechanism for finding regions of interest.

*Rao* et al. [23] proposed a template matching type of model by sliding a template of the desired target to every location in the image and at each location compute salience as some similarity measure between template and local image patch.

*Ma* et al. [33] proposed a user attention model to video contents by incorporating top-down factors into the classical bottom-up framework by extracting semantic cues (e.g., face, speech, and camera motion). First, the video sequence is decomposed into primary elements of basic channels. Next, a set of attention modeling methods generate attention maps separately. Finally, fusion schemes are employed to obtain a comprehensive attention map which may be used as importance ranking or the index of video content. They applied this model to video summarization.

*Rosin* [169] proposed an edge-based scheme (EDS) for saliency detection over grayscale images. First, a Sobel edge detector is applied to the input image. Second, the graylevel edge image is thresholded at multiple levels to produce a set of binary edge images. Third, a distance transform is applied to each of the binary edge images to propagate the edge information. Finally, the gray-level distance transforms are summed to obtain the overall saliency map. This approach has not been successful over color images.

*Garcia-Diaz* et al. [160] introduced the Adaptive Whitening Saliency (AWS) model by adopting the variability in local energy as a measure of saliency estimation. The input image is transformed to *Lab* color space. The luminance (L) channel is decomposed into multi-oriented multi-resolution representation by means of Gabor-like bank of filters. The opponent color components $a$ and $b$ undergo a multi-scale decomposition. By decorrelating the multi-scale responses, extracting from them a local measure of variability, and further performing a local averaging they obtained a unified and efficient measure of saliency. Decorrelation is achieved by applying PCA over a set of multi-scale low level features. Distinctiveness is measured using the Hoteling's $T^2$ statistic.

*Goferman* et al. [46] proposed a context-aware saliency detection model. Salient image regions are detected based on four principles of human attention: 1) Local low-level considerations such as color and contrast, 2) Global considerations which suppress frequently occurring features while maintaining features that deviate from the norm, 3) Visual organization rules which state that visual forms may possess one or several centers of gravity about which the form is organized, and 4) High-level factors, such as human faces. They applied their saliency method to two applications: retargeting and summarization.

Aside from the models discussed so far, there are several other attention models that are relevant to the topic of this review, though they do not explicitly generate saliency maps. Here we mention them briefly.

To overcome the problem of designing the state-space for a complex task, an approach proposed by Sprague and Ballard [109] decomposes a complex temporally-extended task to simple behaviors (also called micro-behaviors), one of which is to attend to obstacles or other objects in the world. This behavior-based approach learns each micro-behavior and uses arbitration to compose these behaviors and solve complex tasks. This complete agent architecture is of interest as it studies the role of attention while it interacts and shares limited resources with other behaviors.

Based on the idea that vision serves action, Jodogne *et al.* [162] introduced an approach for learning action-based image classification known as Reinforcement Learning of Visual Classes (RLVC). RLVC consists of two interleaved learning processes: An RL unit which learns image to action mappings and an image classifier which incrementally learns to distinguish visual classes. RLVC is a feature-based approach in which the entire image is processed to find out whether a specific visual feature exists or not in order to move in a binary decision tree. Inspired by RLVC and U-TREE [163], Borji *et al.* [88] proposed a three-layered approach for interactive object-based attention. Each time the object that is most important to disambiguate appears, a partially unknown state is attended by the biased bottom-up saliency model and recognized. Then the appropriate action for the scene is performed. Some other models in this category are: Triesch *et al.* [97], Mirian *et al.* [100], and Paletta *et al.* [164].

Walker *et al.* [21] built a model based on the idea that humans fixate at those informative points in an image which reduce our overall uncertainty about the visual stimulus - similar to another approach by Lee and Yu [149]. This model is a sequential information maximization approach whereby each fixation is aimed at the most informative image location given the knowledge acquired at each point. A foveated representation is incorporated with reducing resolution as distance increases from the center. Shape histogram edges are used as features.

Lee and Yu [149] proposed that mutual information among the cortical representations of the retinal image, the priors constructed from our long-term visual experience, and a dynamic short-term internal representation constructed from recent saccades, all provide a map for guiding eye navigations. By directing the eyes to locations of maximum complexity in neuronal ensemble responses at each step, the automatic saccadic eye movement system greedily collects information about the external world while modifying the neural representations in the process. This model is close to Najemnik & Geisler's work [20].

To recap, here, we offer a unification of several saliency models from a statistical viewpoint. The first class measures bottom-up saliency as $1/P(x)$ or $logP(x)$ or $E_X[-logP(x)]$ which is the entropy. This includes Torralba and Oliva [92][93], SUN [141], AIM [144], Hou and Zhang [151], and probably Yin Li [171]. Some other methods are equivalent to this but with specific assumptions for $P(x)$. For example, Rosenholtz [191] assume a Gaussian, and Seo and Milanfar [108] assumes that $P(x)$ is a kernel density estimate (with the kernel that appears inside the summation on the denominator of (7)). Next, there is a class of top-down models with the same saliency measure. For example, Elazary and Itti [90] use $logP(x|Y=1)$ (where $Y=1$ means target presence) and assume a Gaussian

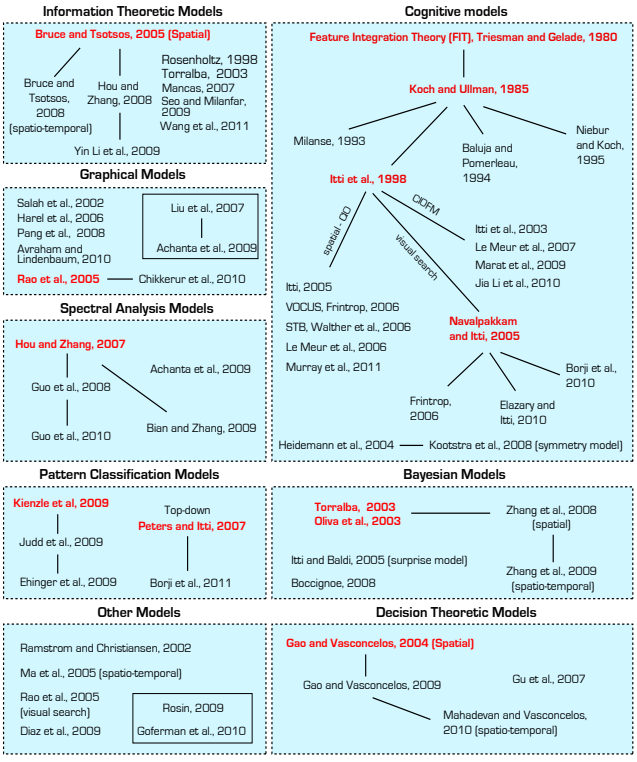



Fig. 6. A hierarchical illustration of described models. Solid rectangles show salient region detection methods.

for $P(x|Y=1)$. SUN can also be seen like this, if you call the first term of (5) a bottom-up component. But, as discussed next, it is probably better to just consider it an approximation to the methods in the third class. The third class includes models that compute posterior probabilities $P(Y=1|X)$ or likelihood ratios $log[P(x|Y=1)/P(x|Y=0)]$. This is the case of discriminant saliency [146][147][215] but also appears in Harel *et al.* [121] (e.g. equation 10) and in Liu *et al.* [43] (if you set the interaction potentials of a CRF to zero, you end up with a computation of the posterior $P(Y=1|X)$ at each location). All these methods model the saliency of each location independently of the others. The final class, graphical models, introduces connections between spatial neighbors. These could be clique potentials in CRFs, edge weights in Harel *et al.* [121], etc.

Fig. 6 shows a hierarchical illustration of models. A summary of attention models and their categorization according to factors mentioned in section 2 is presented in Fig. 7.

## 4 Discussion

There are a number of outstanding issues with attention models that we discuss next.

A big challenge is the degree to which a model agrees with biological findings. Why is such an agreement important? How can we judge whether a model is indeed biologically plausible? While there is no clear answer to these questions in the literature, here we give some hints at their answer. In the context of attention, biologically inspired models have resulted in higher accuracies in some cases. In support of this statement, the Decision theoretic [147][223] and (later) AWS model [160] (and perhaps some other models) are good examples because they explains

| No | Model | Year | f1 | f2 | f3 | f4 | f5 | f6 | f7 | f8 | f9 | f10 | f11 | f12 | f13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bottom-up** (saliency models) | | | | | | | | | | | | | | | |
| 1 | Itti et al. [14] | 1998 | + | - | - | + | - | - | + | f | + | CIO | C | - | - |
| 2 | Privitera & Stark [127] | 2000 | + | - | - | + | - | - | + | f | + | - | O | - | Stark and Choi |
| 3 | Salah et al. [52] | 2002 | + | + | - | + | - | - | + | - | + | O | G | DR | Digit & Face |
| 4 | Itti et al. [119] | 2003 | + | - | + | + | + | + | + | f | + | CIOFM | C | - | - |
| 5 | Torralba [92] | 2003 | - | + | - | + | - | - | + | s | + | CI | B | DR | Torralba et al. |
| 6 | Sun & Fisher [117] | 2003 | + | - | - | + | - | - | + | - | - | CIO | G | - | - |
| 7 | Gao & Vasconcelos [146] | 2004 | - | + | - | + | - | - | + | s | - | DCT | D | DR | Brodatz, Caltech |
| 8 | Ouerhani et al. [210] | 2004 | + | - | - | + | - | - | + | f | + | CIO+Corner | C | CC | Ouerhani |
| 9 | Boccignone & Ferraro [175] | 2004 | + | - | + | - | + | - | + | f | - | Optical Flow | B | - | BEHAVE |
| 10 | Frintrop [50] | 2005 | + | + | + | + | + | + | + | f/s | +/- | CIOM | C | - | - |
| 11 | Itti & Baldi [145] | 2005 | + | - | + | + | + | + | - | f | + | CIOFM | B | KL, AUC | ORIG-MTV |
| 12 | Ma et al. [33] | 2005 | + | - | + | + | - | - | + | f | + | M* | O | - | - |
| 13 | Bruce & Tsotsos [144] | 2006 | + | - | - | + | - | + | + | f | + | DOG, ICA | I | KL, ROC | Bruce and Tsotsos |
| 14 | Navalpakkam & Itti [51] | 2006 | - | + | - | + | - | + | + | s | + | CIO | C | - | - |
| 15 | Zhai & Shah [103] | 2006 | + | - | + | + | + | - | + | f | + | SIFT | O | - | - |
| 16 | Harel et al. [121] | 2006 | + | - | - | + | - | - | + | f | + | IO | G | AUC | Bruce and Tsotsos |
| 17 | Le Meur et al. [41] | 2006 | + | - | - | + | - | - | + | f | + | LM* | C | CC, KL | Le Meur et al. |
| 18 | Walther & Koch [35] | 2006 | + | - | - | + | - | + | + | f | +/- | CIO | C | - | - |
| 19 | Peters & Itti [101] | 2007 | + | + | + | + | + | - | + | i | + | CIOFM | P | KL, NSS | Peters and Itti |
| 20 | Liu et al. [43] | 2007 | + | - | - | + | - | - | + | f | - | Liu* | G | F-measure | Regional |
| 21 | Shic & Scassellati [74] | 2007 | + | - | + | + | + | - | + | f | + | CIOM | C | ROC | Shic and Scassellati |
| 22 | Hou & Zhang [150] | 2007 | + | - | - | + | - | + | + | f | + | FFT, DCT | S | NSS | DB of Hou and Zhang, 2007 |
| 23 | Cerf et al. [167] | 2007 | + | + | - | + | - | + | + | f/s | + | CIO :) | C | AUC | Cerf et al. |
| 24 | Le Meur et al. [138] | 2007 | + | - | + | + | + | - | + | f | + | LM* | C | CC, KL | Le Meur et al. |
| 25 | Mancas [152] | 2007 | + | - | + | + | + | + | + | f | + | CI | I | CC | Le Meur et al. |
| 26 | Guo et al. [156] | 2008 | + | - | - | + | - | - | + | f | + | CIO | D | CC | Self data |
| 27 | Zhang et al. [141] | 2008 | + | - | - | + | - | + | + | f | + | DOG, ICA | B | KL, AUC | Bruce and Tsotsos |
| 28 | Hou & Zhang [151] | 2008 | + | - | + | + | + | - | + | f | + | ICA | I | AUC, KL | Bruce and Tsotsos, ORIG |
| 29 | Pang et al. [102] | 2008 | + | + | + | + | + | - | + | f | + | CIOM | G | NSS | ORIG, Self data |
| 30 | Kootstra et al. [136] | 2008 | + | - | - | + | - | - | + | f | + | Symmetry | C | CC | Kootstra et al. |
| 31 | Ban et al. [172] | 2008 | + | - | + | + | + | - | + | f | + | CIO+SYM | I | - | - |
| 32 | Rajashekar et al. [174] | 2008 | + | - | - | + | - | - | + | f | + | R* | S | CC | Rajashekar et al. |
| 33 | Kienzle et al. [165] | 2009 | + | - | - | + | - | - | + | f | + | I | P | K* | Kienzle et al. |
| 34 | Marat et al. [49] | 2009 | + | - | + | + | + | - | + | f | + | SM* | C | NSS | Marat et al. |
| 35 | Judd et al. [166] | 2009 | + | - | - | + | - | - | + | f | + | J* | P | AUC | Judd et al. |
| 36 | Seo & Milanfar [108] | 2009 | + | - | + | + | + | + | + | f | + | LSK | I | AUC, KL | Bruce and Tsotsos, ORIG |
| 37 | Rosin [169] | 2009 | + | - | - | + | - | - | + | f | + | C+ Edge | O | PR, F-measure | DB of Liu et al, 2007 |
| 38 | Yin Li et al. [171] | 2009 | - | + | + | + | + | + | + | s | + | RGB | S | DR | DB of Hou and Zhang, 2007 |
| 39 | Bian & Zhang [159] | 2009 | + | - | + | + | + | + | + | f | + | FFT | S | AUC | Bruce and Tsotsos |
| 40 | Diaz et al. [160] | 2009 | + | - | - | + | - | + | + | f | + | CIO | O | AUC | Bruce and Tsotsos |
| 41 | Zhang et al. [142] | 2009 | + | - | + | - | + | - | + | f | + | DOG, ICA | B | KL, AUC | Bruce and Tsotsos |
| 42 | Achanta et al. [158] | 2009 | + | - | - | + | - | - | + | f | + | DOG | S | PR | DB of Liu et al, 2007 |
| 43 | Gao et al. [147] | 2009 | + | - | + | + | + | + | + | f | + | CIO | D | AUC | Bruce and Tsotsos |
| 44 | Chikkerur et al. [154] | 2010 | + | + | - | + | - | + | + | f/s | +/- | CIO | B | AUC | Bruce and Tsotsos, Chikkerur |
| 45 | Mahadaven & Vasconcelos [106] | 2010 | + | - | + | - | + | - | + | - | + | I | D | DR, AUC | SVCL background data |
| 46 | Avraham & Lindenbaum [153] | 2010 | + | + | - | + | - | + | + | f/s | +/- | CIO | G | DR, CC | UWGT, Ouerhani et al. |
| 47 | Jia Li et al. [133] | 2010 | - | + | + | + | + | - | + | f | + | CIO | B | AUC | RSD, MTV, ORIG, Peters and Itti |
| 48 | Guo et al. [157] | 2010 | + | - | + | + | + | + | + | f/s | +/- | FFT | S | DR | Self data |
| 49 | Borji et al. [89] | 2010 | - | + | - | + | - | + | + | s | +/- | CIO | O | DR | - |
| 50 | Goeferman et al. [46] | 2010 | + | - | - | + | - | - | + | - | + | C :) | O | AUC | DB of Hou and Zhang, 2007 |
| 51 | Murray et al. [200] | 2011 | + | - | - | + | - | - | + | f | + | CIO | C | AUC, KL | Bruce and Tsotsos, Judd et al. |
| 52 | Wang et al. [201] | 2011 | + | - | - | + | - | - | + | f | + | ICA | I | AUC | Self data |
| **Top-down** (general attention models) | | | | | | | | | | | | | | | |
| 53 | McCallum [163] | 1995 | - | + | - | + | - | + | - | i | + | - | R | - | Self data |
| 54 | Rao et al. [23] | 1995 | - | + | - | + | - | - | + | s | + | CIO | O | - | Self data |
| 55 | Ramstrom & Christiansen [168] | 2002 | - | + | - | + | - | - | + | - | + | CI | O | - | - |
| 56 | Sprague & Ballard [109] | 2003 | - | + | + | - | + | + | + | i | - | S* | R | - | - |
| 57 | Renninger et al. [94] | 2004 | - | + | - | + | - | + | - | s | - | Edgelet | I | DR | Self data |
| 58 | Navalpakkam & Itti [80] | 2005 | - | + | - | + | - | + | + | - | + | CIO | C | - | Self data |
| 59 | Paletta et al. [164] | 2005 | - | + | - | + | - | - | + | - | - | SIFT | R | DR | COIL-20, TSG-20 |
| 60 | Jodogne & Piater [162] | 2007 | - | + | - | + | - | - | + | i | - | SIFT | R | - | - |
| 61 | Butko & Movellan [161] | 2009 | - | + | + | + | + | + | + | s | - | - | R | - | - |
| 62 | Verma & McOwan [214] | 2009 | + | - | - | + | - | + | - | s | - | CIO | O | - | - |
| 63 | Borji et al. [89] | 2010 | - | + | - | + | - | - | + | i | - | CIO | R | - | - |

Fig. 7. Summary of visual attention models. Factors in order are: Bottom-up ($f_1$), Top-down ($f_2$), Spatial (-)/Spatio-temporal (+) ($f_3$), Static ($f_4$), Dynamic ($f_5$), Synthetic ($f_6$) and Natural ($f_7$) stimuli, Task-type ($f_8$), Space-based(+)/Object-based(-) ($f_9$), Features ($f_{10}$), Model type ($f_{11}$), Measures ($f_{12}$), and Used dataset ($f_{13}$). In Task type ($f_8$) column: free-viewing ($f$); target search ($s$); interactive ($i$). In Features ($f_{10}$) column: M* = motion saliency, static saliency, camera motion, object (face) and aural saliency (Speech-music); LM* = contrast sensitivity, perceptual decomposition, visual masking and center-surround interactions; Liu* = center-surround histogram, multi-scale contrast and color spatial-distribution; R* = luminance, contrast, luminance-bandpass, contrast-bandpass; SM* = orientation and motion; J* = CIO, horizontal line, face, people detector, gist, etc; S* = color matching, depth and lines; :) = face. In Model type ($f_{11}$) column, R means that a model is based RL. In Measures ($f_{12}$) column: K* = used Wilcoxon-Mann-Whitney test (The probability that a random chosen target patch receives higher saliency than a randomly chosen negative one); DR means that models have used a measure of detection/classification rate to determine how successful was a model. PR stands for Precision-Recall. In dataset ($f_{13}$) column: Self data means that authors gathered their own data.

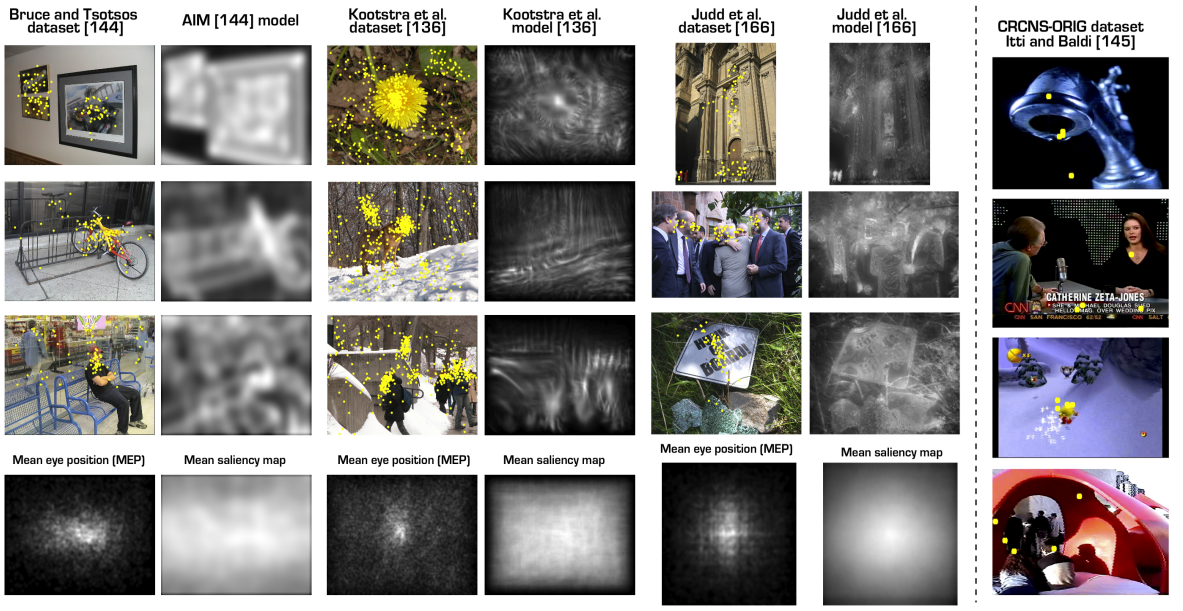



Fig. 8. Sample images from image and video datasets along with eye fixations and predicted attention maps. As could be seen, human and animal body and face, symmetry, and text attract human attention. Fourth row shows that these datasets are highly center-biased mainly because there are some interesting objects at the image center (MEP map). Less center-bias at mean saliency map of models indicates that a Gaussian in average works better than many models.

some basic behavioral data (e.g., nonlinearity against orientation contrast, efficient (parallel) and inefficient (serial) search, orientation and presence-absence asymmetries, and Weber's law [75]) well that has been less explored by other models. These models are among the best in predicting fixations over images and videos [160]. Hence, biological plausibility could be rewarding. We believe that creating a standard set of experiments for judging biological plausibility of models would be a promising direction to take. For some models, prediction of fixations is more important than agreement with biology (e.g., pattern classification vs. cognitive models). These models usually feed features to some classifier - but what type of features or classifiers fall under the realm of biologically inspired techniques? The answer lies in the behavioral validity of each individual feature as well as the classifier (e.g., faces or text, SVM vs. Neural Networks). Note that these problems are not specific to attention modeling and are applicable to other fields in computer vision (e.g., object detection and recognition).

Regarding fair model comparison, results often disagree when using different evaluation metrics. Therefore, a unified comparison framework is required - one that standardizes measures and datasets. We should also discuss the treatment of image borders and its influence on results. For example, KL and NSS measures are corrupted by an edge effect due to variations in handling invalid filter responses at the image borders. Zhang *et al.* [141] studied the impact of varying amounts of edge effects on ROC score over a dummy saliency map (consisting of all ones) and showed that as the border increases, AUC and KL measures increase as well. The dummy saliency map gave an ROC value of 0.5, a four-pixel black border gave 0.62, and an eight-pixel black border map gave 0.73. The same 3 border sizes would yield KL scores of 0, 0.12, and 0.25. Another challenge is handling the center-bias that results from a high density of eye fixations at the image center. Because of this, a trivial Gaussian blob model scores higher than almost all saliency models (see [166]). This can be partially verified from the average eye fixation maps of three popular datasets shown in Fig. 8. Comparing the mean saliency map of models and the fixation distributions, it could be seen that Judd et al. [166] model has higher center-bias due to explicitly using the center feature, which leads to higher eye movement prediction for this model as well. To eliminate the border and center-bias effects, Zhang *et al.* [141] defined an unshuffled AUC metric instead of the uniform AUC metric: for an image, the positive sample set is composed of the fixations of all subjects on that image and the negative set is composed of the union of all fixations across all images - except for the positive samples.

As shown by Figs. 4 and 5 many different eye movement datasets are available, each one recorded in different experimental conditions with different stimuli and tasks. Yet more datasets are needed because the available ones suffer from several drawbacks. Consider that current datasets do not tell us about covert attention mechanisms at all and can only tell us about overt attention (eye tracking). One approximation can compare overt attention shifts to verbal or other reports, whereby reported objects that were not fixated might have been covertly attended to. There is also a lack of multimodal datasets in interactive environments. In this regard, a promising new effort is to create tagged object datasets similar to video LabelMe [188]. Bruce and Tsotsos [144] and ORIG [184] are respectively the most widely used image and video datasets though they are highly center-biased (see Fig. 8). Thus there is a need for standard benchmark datasets as well as rigorous performance measures for attention modeling. Similar efforts have already been started amongst other research communities, such as object recognition (PASCAL challenge), text information retrieval (TREC datasets), and face recognition (e.g., FERET).

The majority of models are bottom-up though it is known

that top-down factors play a major role in directing attention [177]. However, the field of attention modeling lacks principled ways to model top-down attention components as well as the interaction of bottom-up and top-down factors. Feed-forward bottom-up models are general, easy to apply, do not need training, and yield reasonable performance making them good heuristics. On the other hand, top-down definitions usually use feedback and employ learning mechanisms to adapt themselves to specific tasks/environments and stimuli, making them more powerful but more complex to deploy and test (e.g., need to train on large datasets).

Some models need many parameters to be tuned while some others need fewer (e.g., spectral saliency models). Methods such as Gao *et al.* [147], Itti *et al.* [14], Oliva *et al.* [140], and Zhang *et al.* [142]) are based on Gabor or DOG filters and require many design parameters such as the number and type of filters, choice of non-linearities, and normalization schemes. Properly tuning the parameters is important in performance of these types of models.

Fig. 9 presents sample saliency maps of some models discussed in this paper.

## 5 Summary and Conclusion

In this paper, we discussed recent advances in modeling visual attention with an emphasis on bottom-up saliency models. A large body of past research was reviewed and organized in a unified context by qualitatively comparing models over 15 experimental criteria. Advancement in this field could greatly help solving other challenging vision problems such as cluttered scene interpretation and object recognition. In addition, there are many technological applications that can benefit from it. Several factors influencing bottom-up visual attention have been discovered by behavioral researchers and have further inspired the modeling community. However, there are several other factors remaining to be discovered and investigated. Incorporating those additional factors may help to bridge the gap between human inter-observer (a map built from fixations of other subjects over the same stimulus) and prediction accuracy of computational models. With the recent rapid progress, there is hope this may be accessible in the near future.

Most of the previous modeling research has been focused on the bottom-up component of visual attention. While previous efforts are appreciated, the field of visual attention still lacks computational principles for task-driven attention. A promising direction for future research is the development of models that take into account time varying task demands, especially in interactive, complex, and dynamic environments. In addition, there is not yet a principled computational understanding of covert and overt visual attention, which should be clarified in the future. The solutions are beyond the scope of computer vision and require collaboration from the machine learning community.

## Acknowledgments

This work was supported by Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (grant number W911NF-08-1-0360).

The authors would like to thank reviewers for their helpful comments on the paper.

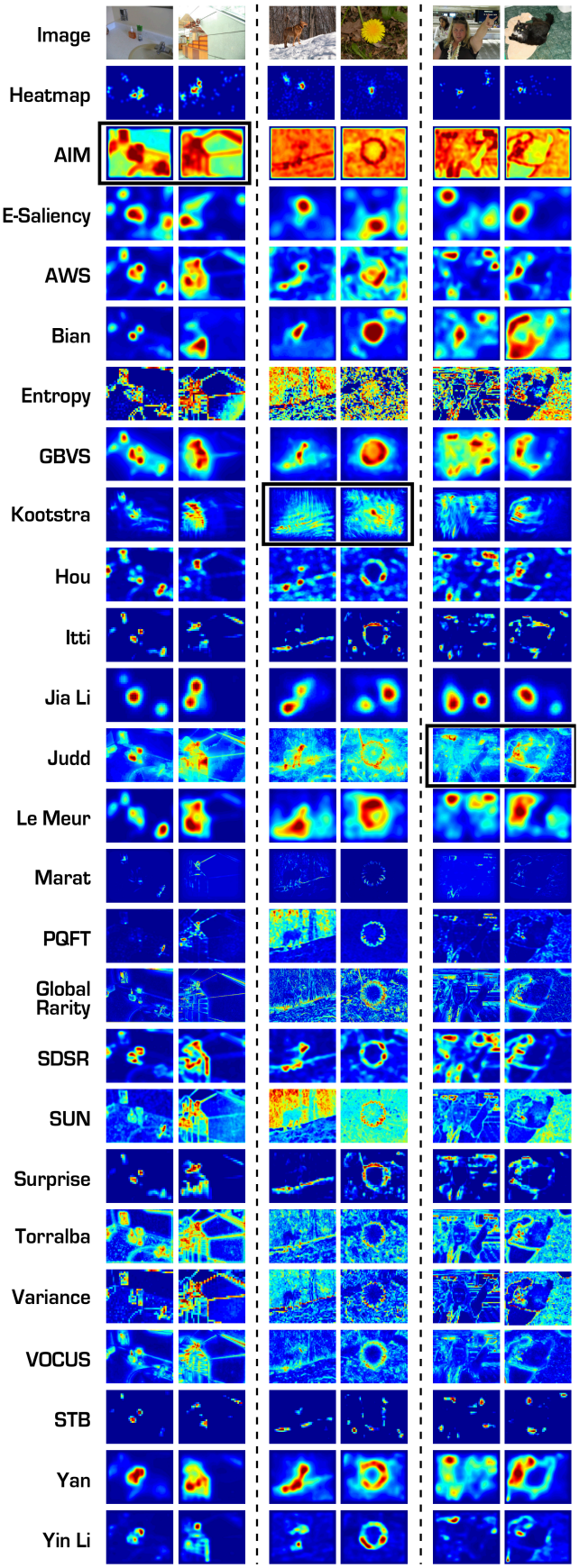



Fig. 9. Sample saliency maps of models over Bruce and Tsotsos (left), Kootstra *et al.* (middle), and Judd et al. datasets. Black rectangles means dataset was first used by that model.

## REFERENCES


[1] K. Koch, J. McLean, R. Segev, M.A. Freed, M.J. Berry, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," *Current Biology*, vol. 25, no. 16(14), pp. 1428-34, 2006.

[2] L. Itti, Models of Bottom-Up and Top-Down Visual Attention, California Institute of Technology, PhD. Thesis, 2000.

[3] D.J. Simons and D.T. Levin, "Failure to Detect Changes to Attended Objects," *Investigative Ophthalmology & Visual Science*, vol. 38, no. 4, pp. 3273, 1997.

[4] R. A. Rensink, "How Much of a Scene is Seen - the Role of Attention in Scene Perception", *Investigative Ophthalmology & Visual Science*, vol. 38, 1997.

[5] D.J. Simons and C.F. Chabris, "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events," *Perception*, vol. 28, no. 9, pp. 1059-1074, 1999.

[6] J.E. Raymond, K.L. Shapiro, and K.M. Arnell, "Temporary Suppression of Visual Processing in an RSVP Task: An Attentional Blink?" *Journal of exp. psych.*, vol. 18, no 3, pp. 849-60, 1992.

[7] S. Treue and J.H.R. Maunsell, "Attentional Modulation of Visual Motion Processing in Cortical Areas MT and MST," *Nature*, vol 382, pp. 539-541, 1996.

[8] S. Frintrop, E. Rome, and H.I. Christensen, "Computational Visual Attention Systems and Their Cognitive Foundations: A Survey," *ACM Trans. Appl. Percept*, vol.7, no. 1. 2010.

[9] A. Rothenstein and J. Tsotsos, "Attention Links Sensing to Recognition," *Image Vision Comput*, vol. 26, pp. 114-126, 2006.

[10] R. Desimone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Annu Rev Neurosci*, vol. 18, pp. 193-222, 1995,

[11] S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone, "Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex," *J. Neurophysiol*, vol. 77, 1997.

[12] C. Bundesen and T. Habekost, "Attention," *In Handbook of Cognition, K. Lamberts and R. Goldstone, Eds.*, 2005.

[13] V. Navalpakkam, C. Koch, A. Rangel, and P. Perona, "Optimal Reward Harvesting in Complex Perceptual Environments," *PNAS*, vol. 107, no. 11, pp. 5232-5237, 2010.

[14] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on PAMI*, vol. 20, no. 11, pp. 1254-1259, 1998.

[15] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling Visual Attention via Selective Tuning," *Artif. Intell.*, vol. 78, no. 1-2, pp. 507-545, 1995.

[16] R. Milanese, Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation, Ph.D. thesis, University of Geneva, Switzerland. 1993.

[17] S. Baluja and D. Pomerleau, "Using a Saliency Map for Active Spatial Selective Attention: Implementation & Initial Results," *NIPS*, pp. 451-458, 1994.

[18] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[19] K. Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, 1998.

[20] J. Najemnik and W.S. Geisler, "Optimal Eye Movement Strategies in Visual Search," *Nature*, no. 434, pp. 387-391, 2005.

[21] L.W. Renninger, J.M. Coughlan, P. Verghese, and J. Malik, "An Information Maximization Model of Eye Movements," *NIPS*, vol. 17, pp. 1121-1128, 2005.

[22] U. Rutishauser and C. Koch, "Probabilistic Modeling of Eye Movement Data During Conjunction Search via Feature-based Attention," *Journal of Vision*, vol. 7, no. 6, 2007.

[23] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, "Eye Movements in Iconic Visual Search," *Vision Res.*, vol. 42, 2002.

[24] A.T. Duchowski, "A Breadth-first Survey of Eye-tracking Applications," *Behav. Res. Methods Instrum Comput*, 2002.

[25] G.E. Legge, T.S. Klitz, and B. Tjan, "Mr. Chips: An Ideal-Observer Model of Reading," *Psychological Review*, 1997.

[26] R.D. Rimey and C.M. Brown, "Controlling Eye Movements with Hidden Markov Models," *IJCV*, vol. 7, no. 1, pp. 47-65, 1991.

[27] S. Treue, "Neural Correlates of Attention in Primate Visual Cortex," *Trends in Neurosciences*, vol. 24, no. 5, pp. 295-300, 2001.

[28] S. Kastner and L.G. Ungerleider, "Mechanisms of Visual Attention in the Human Cortex," *Annu. Rev. Neurosci.*, vol. 23, pp. 315-341, 2000.

[29] E.T. Rolls and G. Deco, "Attention in Natural Scenes: Neurophysiological and Computational Bases," *Neural Networks*, vol. 19, no. 9, pp. 1383-1394, 2006.

[30] G.A. Carpenter and S. Grossberg, "A Massively Parallel Architecture for a Self-organizing Neural Pattern Recognition Machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, no. 1, pp. 54-115, 1987.

[31] N. Ouerhani and H. Hügli, "Real-time Visual Attention on a Massively Parallel SIMD Architecture," *Real-Time Imaging*, vol. 9, no. 3, pp. 189-196, 2003.

[32] Q. Ma, L. Zhang, and B. Wang, "New Strategy for Image and Video Quality Assessment," *J. Electronic Imaging*, vol. 19, 2010.

[33] Y. Ma, X. Hua, L. Lu, and H. Zhang, "A Generic Framework of User Attention Model and Its Application in Video Summarization," *IEEE transactions on multimedia*, vol. 7, no. 5, 2005.

[34] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does Where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric," *ICIP*, vol. 2, pp. 169-172, 2007.

[35] D. Walther and C. Koch, "Modeling Attention to Salient Proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395-1407, 2006.

[36] C. Siagian and L. Itti, "Biologically Inspired Mobile Robot Vision Localization," *IEEE Transactions on Robotics*, 2009.

[37] S. Frintrop and P. Jensfelt, "Attentional Landmarks and Active Gaze Control for Visual SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1054-1065 , 2008.

[38] D. DeCarlo and A. Santella, "Stylization and Abstraction of Photographs," *ACM Trans. on Graphics*, vol. 21, no. 3, 2002.

[39] L. Itti, "Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, 2004.

[40] L. Marchesotti, C. Cifarelli, and G. Csurka, "A Framework for Visual Saliency Detection with Applications to Image Thumbnailing," *ICCV*, 2009.

[41] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE PAMI*, vol. 28, no. 5, pp. 802-817, 2006.

[42] G. Fritz, C. Seifert, L. Paletta and H. Bischof, "Attentive Object Detection Using an Information Theoretic Saliency Measure," *LNCS*, vol. 3368, pp. 29-41, 2005.

[43] T. Liu, J. Sun, N.N Zheng, and H.Y Shum, "Learning to Detect a Salient Object," *CVPR*, 2007.

[44] V. Setlur, R. Raskar, S. Takagi, M. Gleicher, and B. Gooch, "Automatic Image Retargeting, In Mobile and Ubiquitous Multimedia (MUM)," *ACM*, 2005.

[45] C. Chamaret and O. Le Meur, "Attention-based Video reframing: Validation Using Eye-tracking," *ICPR*, 2008.

[46] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," *CVPR*, 2010.

[47] N. Sadaka and L.J. Karam, "Efficient Perceptual Attentive Super-resolution," *IEEE Image Processing (ICIP)*, 2009.

[48] H. Liu, S. Jiang, Q. Huang, and C. Xu, "A Generic Virtual Content Insertion System Based on Visual Attention Analysis," *ACM international conference on Multimedia*, pp. 379-388, 2008.

[49] S. Marat, M. Guironnet, and D. Pellerin, "Video Summarization Using a Visual Attention Model," *EUSIPCO*, 2007.

[50] S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, Springer 2006.

[51] V. Navalpakkam and L. Itti, "An Integrated Model of Top-down and Bottom-up Attention for Optimizing Detection Speed," *CVPR*, 2006.

[52] A. Salah, E. Alpaydin, and L. Akrun, "A Selective Attention-based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420-425, 2002.

[53] S. Frintrop, "General Object Tracking with a Component-based Target Descriptor," *ICRA*, pp. 4531-4536, 2010.

[54] M.S. El-Nasr, T. Vasilakos, C. Rao, and J. Zupko, "Dynamic Intelligent Lighting for Directing Visual Attention in Interactive 3D Scenes," *IEEE Trans. on Comp. Intell. and AI in Games*, 2009.

[55] G. Boccignone, "Nonparametric Bayesian Attentive Video Analysis," *ICPR*, 2008.

[56] G. Boccignone, A. Chianese, V. Moscato, A. Picariello, "Foveated Shot Detection for Video Segmentation," *IEEE Trans. Circuits Syst. Video Techn*. vol. 15, no. 3, pp. 365-377, 2005.

[57] B., Mertsching, M., Bollmann, R., Hoischen, and S. Schmalz, "The neural active vision system." In Handbook of Computer Vision and Applications, Academic Press, 1999.
[58] A. Dankers, N. Barnes, and A. Zelinsky, "A Reactive Vision System: Active-dynamic Saliency," *ICVS*, 2007.
[59] N. Ouerhani, A. Bur, and H. Hügli, "Visual Attention-based Robot Self-localization," *ECMR*, pp. 813, 2005.
[60] S. Baluja, and D. Pomerleau, "Expectation-based Selective Attention for Visual Monitoring and Control of a Robot Vehicle," *Rob. Auton. Syst.*, vol. 22, no. 3-4, pp. 329-344, 1997.
[61] C. Scheier and S. Egner, "Visual Attention in A Mobile Robot," *International Symposium on Industrial Electronics*, pp. 48-53, 1997.
[62] C. Breazeal, "A Context-dependent Attention System for a Social Robot," *IJCAI*, pp. 1146-1151, 1999.
[63] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, "Integrating Context-free and Context-dependent Attentional Mechanisms for Gestural Object Reference," *Mach. Vision Appl.*, vol. 16, no. 1, pp. 64-73, 2004.
[64] G. Heidemann, "Focus-of-attention from Local Color Symmetries," *IEEE Trans PAMI*, vol. 26, no. 7, pp. 817-830, 2004.
[65] A. Belardinelli, Salience Features Selection: Deriving a Model from Human Evidence, Ph.D. thesis, Italy, 2008.
[66] Y. Nagai, "From Bottom-up Visual Attention to Robot Action Learning," *ICDL*, 2009.
[67] C., Muhl, Y., Nagai, and G. Sagerer, "On constructing a communicative space in HRI." *Proceedings of the 30th German Conference on Artificial Intelligence, Springer*, 2007.
[68] T. Liu, S.D. Slotnick, J.T. Serences, and S. Yantis, "Cortical Mechanisms of Feature-based Intentional Control," *Cerebral Cortex*, vol. 13, no. 12, 2003.
[69] B.W. Hong and M. Brady, "A Topographic Representation for Mammogram Segmentation," *LNCS*, vol. 2879, 2003.
[70] N. Parikh, L. Itti, and J. Weiland, "Saliency-based Image Processing for Retinal Prostheses," *Journal of Neural Engineering*, vol 7, no 1, 2010.
[71] O.R. Joubert, D. Fize, G.A. Rousselet, and M. Fabre-Thorpe, "Early Interference of Context Congruence on Object Processing in Rapid Visual Categorization of Natural Scenes," *Journal of Vision*, vol. 8, no. 13, 2008.
[72] H. Li and K.N. Ngan, "Saliency Model-based Face Segmentation and Tracking in Head-and-shoulder Video Sequences," *J. Vis. Commun. Image R.*, vol. 19, pp. 320-333, 2008.
[73] N. Courty and E. Marchand, "Visual Perception based on Salient Features," *IROS*, 2003.
[74] F. Shic and B. Scassellati, "A Behavioral Analysis of Computational Models of Visual Attention," *IJCV*, vol. 73, 2007.
[75] H.C. Nothdurft, "Salience of Feature Contrast," *In Neurobiology of Attention, L. Itti, G. Rees, and J. K. Tsotsos, Eds.*, 2005.
[76] M. Corbetta, and G. L. Shulman, "Control of Goal-directed and Stimulus-driven Attention in the Brain," *Nat. Rev.*, vol. 3, no. 3, pp. 201-215, 2002.
[77] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nat. Rev. Neurosci.*, vol. 2, no. 3, pp. 194-203, 2001.
[78] H.E. Egeth and S. Yantis, "Visual Attention: Control, Representation, and Time Course," *Ann. Rev. Psych.*, vol. 48, 1997.
[79] A.L. Yarbus, Eye-Movements and Vision, Plenum Press, New York, 1967.
[80] V., Navalpakkam and L. Itti, "Modeling the Influence of Task on Attention." *Vision Res.*, vol. 45, no. 2, pp. 205-231, 2005.
[81] A.M. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cognitive Psych.*, vol. 12, pp. 97-136, 1980.
[82] J.M. Wolfe, "Guided Search 4.0: Current Progress with a Model of Visual Search," *In Integrated Models of Cognitive Systems, W. D. Gray, Ed. Oxford University Press, Oxford, UK*, 2007.
[83] G.J. Zelinsky, "A Theory of Eye Movements During Target Acquisition," *Psychological Review*, vol. 115, no. 4, 787-835, 2008.
[84] W. Einhauser, M. Spain, and P. Perona, "Objects Predict Fixations Better Than Early Saliency," *Journal of Vision*, 2008.
[85] M. Pomplun, "Saccadic Selectivity in Complex Visual Search Displays," *Vision Research*, vol. 46, pp. 1886-1900, 2006.
[86] A. Hwang and M. Pomplun, "A Model of Top-down Control of Attention During Visual Search in Real-world Scenes," *Journal of vision*, vol. 8, no. 6, pp. 2008.
[87] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modeling Search for People in 900 Scenes: A Combined Source Model of Eye Guidance," *Visual Cognition*, vol. 17, 2009.
[88] A. Borji, M.N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online Learning of Task-driven Object-based Visual Attention Control," *Image. Vision Comput.*, vol. 28, pp. 1130-1145, 2010.
[89] A. Borji, M.N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive Learning of Top-down Modulation for Attentional Control," *Machine Vision and Applications*, vol. 22, 2011.
[90] L. Elazary and L. Itti, "A Bayesian Model for Efficient Visual Search and Recognition," *Vision Research*, vol. 50, 2010.
[91] M.M. Chun and Y. Jiang, "Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention," *Cognitive Psychology*, vol. 36, pp. 28-71, 1998.
[92] A. Torralba, "Modeling Global Scene Factors in Attention," *Journal of Optical Society of America*, vol. 20, no. 7, 2003.
[93] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope," *International Journal in Computer Vision*, vol. 42, pp. 145-175, 2001.
[94] L.W. Renninger and J. Malik, "When is Scene Recognition Just Texture Recognition?" *Vis. Res.*, vol. 44, pp. 2301-2311, 2004.
[95] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE PAMI*, vol. 29, no. 2, pp. 300-312, 2007.
[96] M. Viswanathan, C. Siagian, and L. Itti, Vision Science Symposium (VSS), 2007.
[97] J. Triesch, D.H. Ballard, M.M. Hayhoe, and B.T. Sullivan, "What You See Is What You Need," *Journal of Vision*, 2003.
[98] M.I. Posner, Orienting of Attention, Q. J. Exp. Psych. vol. 32, pp. 3-25, 1980.
[99] M. Hayhoe and D. Ballard, "Eye Movements in Natural Behavior," *Trends in Cognitive Sciences*, vol. 9, pp. 188-194, 2005.
[100] M.S. Mirian, M. N. Ahmadabadi, B.N. Araabi, R. R. Siegwart, "Learning Active Fusion of Multiple Experts' Decisions: An Attention-based Approach," Neural Computation, 2011.
[101] R.J. Peters and L. Itti, "Beyond Bottom-up: Incorporating Task-dependent Influences Into a Computational Model of Spatial Attention," *CVPR*, 2007.
[102] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network," *ICME*, 2008.
[103] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues',' *ACM International Conference on Multimedia*, 2006.
[104] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modeling Spatio-temporal Saliency to Predict Gaze Direction for Short Videos," *IJCV*, 2009.
[105] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes." *IEEE PAMI*, vol. 32, no. 1, 2010.
[106] V. Mahadevan and N. Vasconcelos, "Saliency Based Discriminant Tracking." *In IEEE (CVPR)*, 2009.
[107] N. Jacobson, Y-L. Lee, V. Mahadevan, N. Vasconcelos and T.Q. Nguyen, "A Novel Approach to FRUC using Discriminant Saliency and Frame Segmentation." *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2924-2934, 2010.
[108] H.J. Seo and P. Milanfar, "Static and Space-time Visual Saliency Detection by Self-Resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1-27, 2009.
[109] N. Sprague and D.H. Ballard, "Eye Movements for Reward Maximization," *NIPS*, 2003.
[110] http://tcts.fpms.ac.be/ mousetrack/
[111] J. Bisley and M. Goldberg, "Neuronal Activity in The Lateral Intraparietal Area and Spatial Attention," *Science*, 2003.
[112] J. Duncan, "Selective Attention and The Organization of Visual Information," *J. Exp. Psych.*, vol. 113, pp. 501-517, 1984.
[113] B.J. Scholl, "Objects and Attention: The State of The Art," *Cognition*, vol. 80, pp. 1-46. 2001.
[114] Z. W. Pylyshyn and R. W. Storm, "Tracking Multiple Independent Targets: Evidence for a Parallel Tracking Mechanism," *Spatial Vision*, vol. 3, pp. 179-197, 1988.
[115] E. Awh and H. Pashler, "Evidence For Split Attentional Foci," *J. Exp.Psych. Hum. Percept. Perform.*, vol. 26, pp. 834-846, 2000.
[116] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "LabelMe: A Database and Web-based Tool for Image Annotation," *IJCV*, vol. 77, no. 1-3, pp. 157-173, 2008.
[117] Y. Sun and R. Fisher, "Object-based Visual Attention for Computer Vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77-123, 2003.
[118] J.M. Wolfe and T.S. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nat. Rev. Neurosci.*, vol. 5, pp. 1-7, 2004.

[119] L. Itti, N. Dhavale, and F. Pighin, "Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention," *SPIE*, vol. 5200, pp. 64-78, 2003.
[120] R. Rae, Gestikbasierte Mensch-Maschine-Kommunikation auf der Grundlage visueller Aufmerksamkeit und Adaptivität. Ph.D. thesis, Universität Bielefeld, Germany, 2000.
[121] J. Harel, C. Koch, and P. Perona, "Graph-based Visual Saliency," *NIPS*, vol. 19, pp. 545-552, 2006.
[122] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video," *ICCV*, 2005.
[123] B.W. Tatler, "The Central Fixation Bias in Scene Viewing: Selecting an Optimal Viewing Position Independently of Motor Bases and Image Feature Distributions," *J. Vision*, 2007.
[124] R. Milanese, Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation, Ph.D. thesis, University of Geneva, Switzerland, 1993.
[125] F.H. Hamker, "The Emergence of Attention by Poulation-based Inference and Its Role in Distributed Processing and Cognitive Control of Vision," *J. Comput. Vision Image Understanding*, vol. 100, no. 1-2, pp. 64106. 2005.
[126] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt Visual Attention For a Humanoid Robot," *IROS*, 2001.
[127] C.M. Privitera and L.W. Stark, "Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations," *IEEE PAMI*, vol. 22, no. 9, pp. 970-982, 2000.
[128] K. Lee, H. Buxton, and J. Feng, "Selective Attention for Cue-guided Search Using a Spiking Neural Network," *WAPCV*, pp. 5562, 2003.
[129] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int. J. Comput. Vision*, vol. 45, no. 2, pp. 83-105, 2001.
[130] A. Maki, P. Nordlund, and J.O. Eklundh, "Attentional Scene Segmentation: Integrating Depth and Motion," *Comput. Vision Image Understanding*, vol. 78, no. 3, pp. 351-373, 2000.
[131] D. Parkhurst, K. Law, and E. Niebur, "Modeling the Role of Salience in The Allocation of Overt Visual Attention," *Vision Res*. vol. 42, no. 1, pp. 107-123, 2002.
[132] T.S. Horowitz, and J.M. Wolfe, "Visual Search Has No Memory," *Nature*, vol. 394, pp. 575-577, 1998.
[133] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video," *IJCV*, 2010.
[134] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of Bottom-up Gaze Allocation in Natural Images," *Vis. Res.*, 2005.
[135] M. Land and M. Hayhoe, "In What Ways Do Eye Movements Contribute to Everyday Activities?" *Vis. Res.*, vol. 41, 2001.
[136] G. Kootstra, A. Nederveen, and B. de Boer, "Paying Attention to Symmetry," *BMVC*, pp. 1115-1125, 2008.
[137] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free Attentional Operators: The Generalized Symmetry Transform," *Int. Journal of Computer Vision*, vol. 14, no. 2, pp. 119-130, 1995.
[138] O. Le Meur, P. Le Callet and D. Barba, "Predicting Visual Fixations on Video Based on Low-level Visual Features," *Vision Research*, vol. 47/19, pp. 2483-2498, 2007.
[139] D.D. Salvucci, "An Integrated Model of Eye Movements and Visual Encoding," *Cognitive Systems Research*, vol. 1, 2001.
[140] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson, "Top-down Control of Visual Attention in Object Detection," *ICIP*, pp. 253-256, 2003.
[141] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. of Vision*, vol. 8(7), no. 32, pp. 1-20, 2008.
[142] L. Zhang, M.H. Tong, and G.W. Cottrell, "SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes," *In Thirty-first Annual Cognitive Science Society Conference*, 2009.
[143] N.D.B. Bruce and J.K. Tsotsos, "Spatiotemporal Saliency: Towards a Hierarchical Representation of Visual Saliency," WAPCV, 2008.
[144] N.D.B. Bruce, J.K. Tsotsos, "Saliency Based on Information Maximization," *NIPS*, 2005.
[145] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *NIPS*, 2005.
[146] D. Gao and N. Vasconcelos, "Discriminant Saliency for Visual Recognition from Cluttered Scenes," *NIPS*, 2004.
[147] D. Gao, S. Han and N. Vasconcelos, "Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition." IEEE Trans. PAMI. vol. 31, no. 6, 2009.
[148] E. Gu, J. Wang, and N.I. Badler, "Generating Sequence of Eye Fixations Using Decision-Theoretic Attention Model," *WAPCV*, pp. 277-29, 2007.
[149] T.S. Lee and S. Yu, "An Information-theoretic Framework for Understanding Saccadic Behaviors," *NIPS*, 2000.
[150] X. Hou and L. Zhang. "Saliency Detection: A Spectral Residual Approach," *CVPR*, 2007.
[151] X. Hou and L. Zhang. "Dynamic Visual Attention: Searching for Coding Length Increments," *NIPS*, 2008.
[152] M. Mancas, Computational Attention: Modelisation and Application to Audio and Image Processing, PhD. thesis, 2007.
[153] T. Avraham, M. Lindenbaum, "Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling," *IEEE PAMI*, vol. 32, no. 4, pp. 693-708, 2010.
[154] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and Where: A Bayesian Inference Theory of Visual Attention," Vision Research, 2010.
[155] P. Verghese, "Visual Search and Attention: A Signal Detection Theory Approach," *Neuron*, vol. 31, pp. 523-535, 2001.
[156] C. Guo, Q. Ma, and L. Zhang, "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *CVPR*, 2008.
[157] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185-198, 2010.
[158] R. Achanta, S.S. Hemami, F.J. Estrada, and S. Süsstrunk, "Frequency-tuned Salient Region Detection," *CVPR*, 2009.
[159] P. Bian and L. Zhang, "Biological Plausibility of Spectral Domain Approach for Spatiotemporal Visual Saliency," *LNCS*, vol. 5506, pp. 251-258, 2009.
[160] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and Distinctiveness Provide With Human-Like Saliency," *ACIVS*, vol. 5807, 2009.
[161] N.J. Butko and J.R. Movellan, "Optimal Scanning for Faster Object Detection," *CVPR*, 2009.
[162] S. Jodogne and J. Piater, "Closed-Loop Learning of Visual Control Policies," *Journal of Artificial Intelligence Research*, vol. 28, pp. 349-391, 2007.
[163] R. McCallum, Reinforcement Learning with Selective Perception and Hidden State, Ph.D thesis, 1996.
[164] L. Paletta, G. Fritz, and C. Seifert, "Q-learning of Sequential Attention for Visual Object Recognition From Informative Local Descriptors," *ICML*, pp. 649-656, 2005.
[165] W. Kienzle, M.O. Franz, B. Schölkopf, and F.A. Wichmann, "Center-surround Patterns Emerge as Optimal Predictors for Human Saccade Targets," *Journal of Vision*, vol. 9, 2009.
[166] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *ICCV*, 2009.
[167] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting Human Gaze Using Low-level Saliency Combined With Face Detection," *NIPS*, 2007.
[168] O. Ramström and H.I. Christensen, "Visual Attention Using Game Theory," *Biologically Motivated Computer Vision Conference*, pp. 462-471, 2002.
[169] P.L. Rosin, "A Simple Method for Detecting Salient Regions," *Pattern Recognition*, vol. 42, no. 11, pp. 2363-2371, 2009.
[170] Z. Li, "A Saliency Map in Primary Visual Cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9-16, 2002.
[171] Y. Li, Y. Zhou, J. Yan, and J. Yang, "Visual Saliency Based on Conditional Entropy," *ACCV*, 2009.
[172] S.W Ban, I. Lee, and M. Lee, "Dynamic Visual Selective Attention Model," *Neurocomputing*, vol. 71, no. 4-6, 2008.
[173] M. T. López, M. A. Fernńdez, A. Fernández-Caballero, J. Mira, A.E. Delgado, "Dynamic Visual Attention Model in Image Sequences," *Image and Vision Computing*, vol. 25, 2007.
[174] U. Rajashekar, I. van der Linde, A.C. Bovik, and L. K., Cormack, "GAFFE: A Gaze-Attentive Fixation Finding Engine," *IEEE Trans. on Image Processing*, vol. 17, no. 4, pp. 564-573, 2008.
[175] G. Boccignone and M. Ferraro, "Modeling gaze shift as a constrained random walk", Physica A, vol. 331, 2004.
[176] M. C. potter, "Meaning in Visual Scenes," *Science*, vol. 187, pp. 965-966, 1975.
[177] J. M. Henderson and A. Hollingworth, "High-level Scene Perception," *Ann. Rev. of Psychology*, vol. 50, pp. 243-271, 1999.
[178] R.A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, vol. 7, pp. 17-42, 2000.

[179] J. Bailenson and N. Yee, "Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments," *Psychological Science*, vol. 16, pp. 814-819, 2005.
[180] M. Sodhi , B. Reimer, J.L. Cohen, E. Vastenburg, R. Kaars, and S. Kirschenbaum, "On-road Driver Eye Movement Tracking Using Head-mounted Devices," *Proceedings of the Symposium on Eye tracking Research & Applications*, 2002.
[181] J.H. Reynolds and D.J. Heeger, "The Normalization Model of Attention," *Neuron*, vol. 61, no. 2, pp. 168-185, 2009.
[182] S. Engmann, B.M. Hart, T. Sieren, S. Onat, P. König, and W. Einhäuser, "Saliency on a Natural Scene Background: Effects of Color and Luminance Contrast Add Linearly," *Attention, Perception & Psychophysics*, vol. 71, no. 6, pp. 1337-1352, 2009.
[183] A. Reeves and G. Sperling, "Attention Gating in Short-term Visual Memory," *Psych. Review*, vol. 93, no. 2, pp. 180-206, 1986.
[184] L. Itti, "Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093-1123, 2005.
[185] D. Gao, V. Mahadevan and N. Vasconcelos, "On the Plausibility of the Discriminant Center-surround Hypothesis for Visual Saliency," *Journal of Vision*, vol. 8, no. (7):13, pp. 1-18, 2008.
[186] J. Yan, J. Liu, Y. Li, and Y. Liu, "Visual Saliency via Sparsity Rank Decomposition," *ICIP*, 2010.
[187] http://www.its.caltech.edu/∼xhou/
[188] J. Yuen, B.C. Russell, C. Liu, and A. Torralba, "LabelMe Video: Building a Video Database with Human Annotations," *ICCV*, 2009.
[189] R. Rosenholtz, Y. Li, and L. Nakano, Measuring Visual Clutter. Journal of Vision, vol. 7(2), no. 17, pp. 1-22, 2007.
[190] R. Rosenholtz, A. Dorai, and R. Freeman, "Do Predictions of Visual Perception Aid Design?" *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, 2011.
[191] R. Rosenholtz, "A Simple Saliency Model Predicts a Number of Motion Popout Phenomena," *Vis. Res.*, vol. 39, 1999.
[192] X. Hou, J. Harel, and Christof Koch, "Image Signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.
[193] R. Rosenholtz, A.L. Nagy, and N. R. Bell, "The Effect of Background Color on Asymmetries in Color Search", *Journal of Vision*, vol. 4, no. 3, pp. 224-240, 2004.
[194] http://alpern.mit.edu/saliency/
[195] D. Green and J. Swets, Signal Detection Theory and Psychophysics, New York: John Wiley, 1966.
[196] T. Jost, N. Ouerhani, R. von Wartburg, R. Mäuri, and H. Häugli, "Assessing the Contribution of Color in Visual Attention," *Computer Vision and Image Understanding*, vol. 100, 2005.
[197] U. Rajashekar, A.C. Bovik, and L.K. Cormack, "Visual Search in Noise: Revealing the Influence of Structural Cues by Gaze-contingent Classification Image Analysis," *J. Vision*, 2006.
[198] S.A. Brandt and L.W. Stark, Spontaneous Eye Movements during Visual Imagery Reflect the Content of the Visual Scene," *Journal of Cognitive Neuroscience*, vol. 9, no. 27-38,1997.
[199] A.D. Hwang, H.C. Wang, and M. Pomplun, "Semantic Guidance of Eye Movements in Real-world Scenes," *Vis. Res*, 2011.
[200] N. Murray, M. Vanrell, X. Otazu, and C. Alejandro Parraga, "Saliency Estimation Using a Non-Parametric Low-Level Vision Model," *CVPR*, 2011.
[201] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating Human Saccadic Scanpaths on Natural Images," *CVPR*, 2011.
[202] R.L Canosa, "Real-World Vision: Selective Perception and Task," *ACM Transactions on Applied Perception*, vol. 6, no. 2, 2009.
[203] M.S. Peterson, A.F. Kramer, and D.E. Irwin, "Covert Shifts of Attention Precede Involuntary Eye Movements," Perception & Psychophysics, vol. 66, pp. 398-405, 2004.
[204] F. Baluch and L. Itti, "Mechanisms of Top-Down Attention," *Trends in Neuroscience*, vol. 34, no. 4, 2011.
[205] J. Hayes and A. Efros, "Scene Completion Using Millions of Photographs", *SIGGRAPH*, 2007.
[206] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. PAMI*, vol. 32, no. 9, 2010.
[207] A.K. Mishra and Y. Aloimonos, "Active Segmentation," *International Journal of Humanoid Robotics*, vol 6., pp 361-386, 2009.
[208] B. Suh, H. Lingm, B.B. Bederson, and D.W. Jacobs. "Automatic Thumbnail Cropping and Its Effectiveness," In *UIST*, pp. 95-104, 2003
[209] S. Mitri, S. Frintrop, K. Pervolz, H. Surmann, and A. Nuchter, "Robust Object Detection at Regions of Interest with an Application in Ball Recognition," *ICRA*, pp. 126-131, 2005.
[210] N. Ouerhani, R. von Wartburg, H. Hugli, and R.M. Muri, "Empirical Validation of Saliency-based Model of Visual Attention," *Electronic Letters on Computer Vision and Image Analysis*, vol. 3, no. 1, pp. 13-24, 2003.
[211] L.W. Stark and Y. Choi, "Experimental Metaphysics: The Scanpath as an Epistemological Mechanism," *Visual Attention and Cognition.* pp. 3-69, 1996.
[212] P. Reinagel and A. Zador, "Natural Scenes at the Center of Gaze," *Network*, vol. 10, pp. 341-50, 1999.
[213] U. Engelke, H.J. Zepernick, and A. Maeder, "Visual Attention Modeling: Region-of-interest Versus Fixation Patterns", *Picture Coding Symposium (PCS)*, 2009.
[214] M. Verma and P.W. McOwana, "Generating Customised Experimental Stimuli for Visual Search Using Genetic Algorithms Shows Evidence For a Continuum of Search Efficiency," *Vision Research*, vol. 49, no. 3, pp. 374-382, 2009.
[215] S. Han and N. Vasconcelos, "Biologically Plausible Saliency Mechanisms Improve Feedforward Object Recognition", *Vision Research*, vol. 50, no. 22, pp. 2295-2307, 2010.
[216] D. Ballard, M. Hayhoe, J. Pelz, "Memory Representations in Natural Tasks," *J. of Cognitive Neuroscience*, vol. 7, no. 1, 1995.
[217] R. Rao, "Bayesian Inference and Attentional Modulation in the Visual Cortex," *NeuroReport*, vol. 16, no. 16, 2005.
[218] A.Borji, D.N. Sihite, and L. Itti, "Computational Modeling of Top-down Visual Attention in Interactive Environments," *BMVC*, 2011.
[219] E. Niebur and C.Koch, "Control of Selective Visual Attention: Modeling the Where Pathway," *NIPS*, pp. 802-808, 1995.
[220] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *IJCV*, vol. 57, no. 2, pp. 137-154, 2004.
[221] W. Kienzle, B. Schø"lkopf, F.A. Wichmann, M.O. Franz, "How to Find Interesting Locations in Video: A Spatiotemporal Interest Point Detector Learned from Human Eye Movements," *DAGM-Symposium*, pp. 405-414, 2007.
[222] J. Wang, J. Sun, L. Quan, X. Tang, and H.Y Shum, "Picture Collage," CVPR, 2006.
[223] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, pp. 239-271, 2009.
[224] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, pp. 1484-1525, 2011.

**Ali Borji** received the BS and MS degrees in computer engineering from Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He did his Ph.D. in cognitive neurosciences at Institute for Studies in Fundamental Sciences (IPM) in Tehran, Iran, 2009. He is currently a postdoctoral scholar at iLab, University of Southern California, Los Angeles, CA. His research interests include: visual attention, visual search, machine learning, robotics, neurosciences and biologically plausible vision models.

**Laurent Itti** received his M.S. degree in Image Processing from the Ecole Nationale Superiere des Télecommunications in Paris in 1994 and his Ph.D. in Computation and Neural Systems from Caltech in 2000. He is now and associate professor of Computer Science, Psychology and Neurosciences at the University of Southern California. Dr Itti's research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, gist, saliency, and surprise, with technological applications to video compression, target detection, and robotics.

Vision Research 50 (2010) 1338–1352

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

# A Bayesian model for efficient visual search and recognition

Lior Elazary [a,*], Laurent Itti [b]

[a] Department of Computer Science, University of Southern California, Los Angeles, CA 90089-2520, USA
[b] Department of Computer Science and Neuroscience Graduate Program, University of Southern California, Los Angeles, CA 90089-2520, USA

ARTICLE INFO

*Article history:*
Received 24 July 2009
Received in revised form 13 November 2009

*Keywords:*
Recognition
Search
Attention
Feature
Scene analysis

ABSTRACT

Humans employ interacting bottom-up and top-down processes to significantly speed up search and recognition of particular targets. We describe a new model of attention guidance for efficient and scalable first-stage search and recognition with many objects (117,174 images of 1147 objects were tested, and 40 satellite images). Performance for recognition is on par or better than SIFT and HMAX, while being, respectively, 1500 and 279 times faster. The model is also used for top-down guided search, finding a desired object in a $5 \times 5$ search array within four attempts, and improving performance for finding houses in satellite images.

© 2010 Elsevier Ltd. All rights reserved.

## 1. Introduction

Attempting to search for and recognize particular known objects in a scene can be extremely complex when one has to consider all possible views an object can take. Humans employ attention to try to limit the amount of information that needs to be processed in order to speed up search and recognition (we rarely look at the sky when searching for our car). Previous research has shown that visual search tasks can be performed faster when one knows the exact target in visual space, as opposed to only a semantic description of the target (Wolfe, 1998). Therefore, humans use cues from the target image to help facilitate search. One can also consider implementing attention in the feature domain when searching through a large dataset for a particular object. For example, if we wish to search for a green bottle, we could bias the visual system so that green vertical edges would be perceived faster than other features (since bottles are often upright). This would allow us to focus a more complex recognition onto only the locations in the search scene that contain green vertical edges, which would speed up the search significantly. Likewise, during recognition, that green vertical edge may be useful to quickly narrow down onto a smaller set of possible recognition candidates. The use of various features in this manner can help sift through very large object datasets when attempting to recognize objects (consider the large number of objects that an adult human can identify). Lastly, it has been shown by Tsotsos (1991) that knowing the features of a target reduces the complexity of visual search from NP complete to linear. These findings suggest that humans employ various heuristics to improve the tractability of performing search and recognition. In this paper, we develop a model which explores the use of biologically plausible attentional heuristics to speed up search and recognition.

It is well known that the search and recognition behavior in humans can be explained through the combination of bottom-up information from the incoming visual scene (Itti & Koch, 2001; Theeuwes, 1995), and of top-down information from the visual knowledge of the target and the scene (Moran & Desimone, 1985; Motter, 1994; Treue & Trujillo, 1999; Wolfe et al., 2004; Krummenacher, Muller, Reimann, & Heller, 2001; Theeuwes, 1994; Hayhoe & Ballard, 2005). However, the exact interaction between the two processes still remains elusive, which has made it difficult to develop machine vision systems exploiting both bottom-up and top-down information.

There have been at least three major theories on mechanisms of integration between bottom-up and top-down vision occurring in the visual cortex. The first is Feature Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990), in which several low-level visual features are processed over the entire visual field in separate neuronal maps (called feature maps), and then combined to form a master map that guides attention. If the target can be defined by a set of primitive feature maps (e.g., it has a distinct color, orientation, intensity), then these maps can be biased using such top-down information to elicit the target location. However, if the target is defined only by some conjunctions of these primitive feature maps (e.g., a unique combination of color and orientation), then a serial search is required to find the target, since a unique signature of the target cannot be obtained from the separate feature maps alone. In

* Corresponding author.
*E-mail addresses:* elazary@usc.edu (L. Elazary), itti@usc.edu (L. Itti).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.01.002

contrast, the Guided Search method proposed by Wolfe (1994) creates a master activation map where top-down knowledge is used to weigh the relative contributions of bottom-up feature maps to emphasize both features (e.g., a red color) and locations (e.g., the top-left corner of the image) likely to characterize the target. The model then uses the combination of these weighted maps to shift attention towards the most promising locations. Lastly, the Biased Competition Model proposed by Desimone and Duncan (1995) involves competition between visual stimuli at each stage of processing, which is influenced by top-down modulation. In this model, attention biases the response of a local feature detector when two stimuli are simultaneously exciting it (i.e., are presented within the receptive field of the same visual neuron). The response is biased in the direction of the attended feature in a different location. In all these models, choosing the correct feature maps to use in visual search, as well as deciding how exactly to influence these maps with top-down information, is crucial to search performance.

Previous models such as Feature Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990), Guided Search (Wolfe, 1994), Biased Competition Model (Desimone & Duncan, 1995) and Optimal Gains (Navalpakkam & Itti, 2006a, 2007) have largely concentrated on biasing the feature maps in a global way to facilitate efficient search (Fig. 1). For example, changing gains (or weights) over whole maps has been proposed and implemented in Wolfe (1994), Navalpakkam and Itti (2006a) and Treue and Trujillo (1999). However, simply setting feature gains globally may not always accelerate search for a target object, especially for maps that code for features shared by the target and many distractors. Furthermore, previous models have concentrated on determining the values of these gains from the objects so as to guide search towards them, but most have not shown how they can be used for object recognition. In this work a common representational framework is used for learning how to bias towards desired targets and for recognizing these targets when they are found. Thus allowing the same top-down signals or parameters used for attention biasing, to also be used for recognition.

One of the previous proposals to compute the gain or weight of particular feature maps is to base the values on the signal to noise ratio, defined as the ratio of a detector's response to the target relative to a distractor. Namely, this approach proposed that the relative weights of feature maps should be modulated top-down in proportion to each map's ability to distinguish the target from the distractors (Navalpakkam & Itti, 2006a; Navalpakkam & Itti, 2007). One shortcoming of such an approach is that, if the detectors in a given feature map respond to both the target and the distractors equally, then no change in gain will take place (Fig. 2a), which would not contribute to improvement of search speed. Moreover, if a feature detector responds more strongly to a distractor object than to the target, a reduction in gain of this map would occur, which could end up turning off this map completely. As a result, only the feature maps that can uniquely distinguish the object being searched for are amplified. Nonetheless, if a target object contains a weak red feature among strong red distractors, the weak red signal could in principle be used to find the object by guiding attention towards locations where feature detectors report low red values. Even if the feature maps are divided into sub-bands with finer granularity (Fig. 2a and b) as proposed in Navalpakkam and Itti (2006b), one can always design search arrays in which one band can code for both the target and distractors, leading to a failed discrimination.

There have also been many contributions to object recognition and search in the computer vision literature. These contributions often concentrate on two aspects of the problem: developing methods to extract features from images, and creating algorithms to classify these features. Some of the research has also independently been focused on searching for objects once particular features have been learned. For example, simple template matching (Gonzalez & Wintz, 1987; Horn, 1986; Pratt, 1991; MacLean & Tsotsos, 2008) or back-projection approaches (Bradski, 1998; Comaniciu & Meer, 1997) use some knowledge (a template or histogram) to check every possible location in the image for a good match. These techniques often fail when the object's pose or

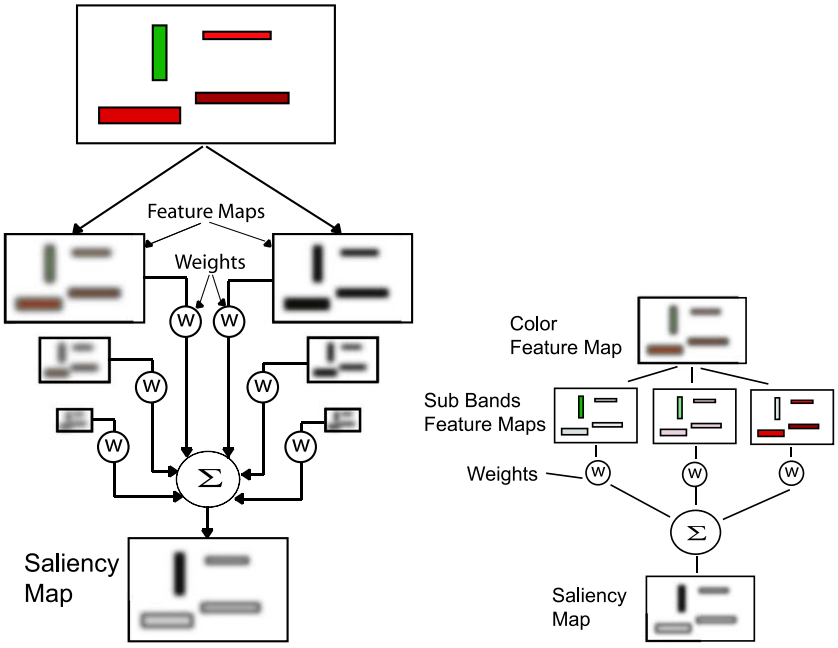



**Fig. 1.** Example search in previous models such as Feature Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990), Guided Search (Wolfe, 1994), Biased Competition Model (Desimone & Duncan, 1995) and Optimal Gains (Navalpakkam & Itti, 2006a, 2007). Left image shows the basic scheme of computing a saliency map from the weighted sum of various feature maps with varying scales (intensity, color, orientation, etc.). Biasing the saliency map towards a particular feature or scale can be achieved by changing the relative weights ($w$) between the feature maps. Right image shows how greater granularity in biasing can be accomplished by splitting a particular feature map into multiple sub-bands. Ultimately, both models fail to provide fine granularity in biasing for specific features (see text for explanation).

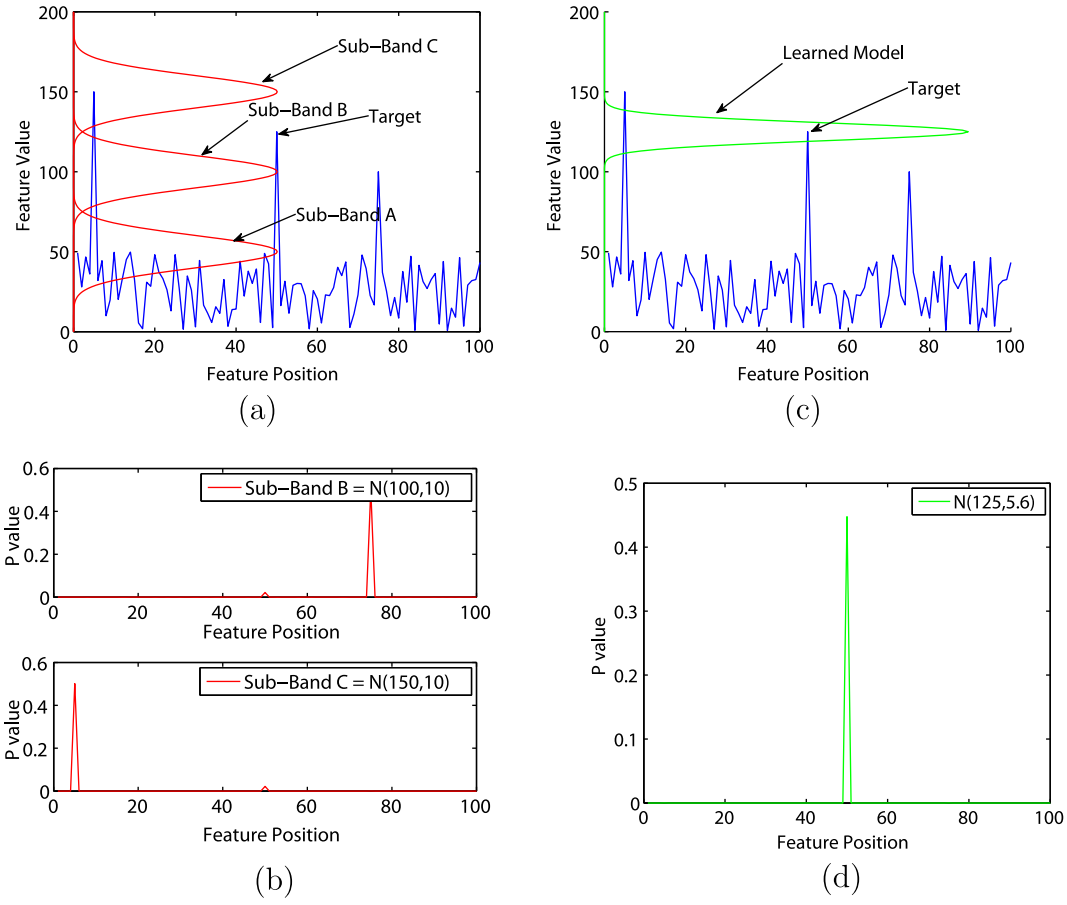



**Fig. 2.** An example of biasing using feature bands (a,b) and a likelihood model (c,d). In both cases the target (at spatial position 50 in a 1D slice of a feature map) has a feature value of 125. (a) Shows how three sub-bands with mean feature responses at 50, 100, 150 and standard deviation of 10 will split the feature space. (b) Shows the ambiguity in the response of sub-bands B and C when searching for the target, whereby each sub-band responds more vigorously to a distractor than to the target (sub-band A does not respond at all to the target and is not shown). As a result, changing the weight of any one of them will not yield a higher response for the target. (c) Shows how knowing the model of the target can give the granularity needed to find the target, while (d) shows the response from the learned model.

illumination is changed. To speed up search, an attentional framework proposed by Bonaiuto and Itti (2005) uses a bottom-up saliency map to rapidly eliminate locations in scenes which are unlikely to contain interesting objects. Although they reported faster results in their searches, the system lacked a method for exploiting top-down knowledge about the search target's features. Obdrzalek and Matas (2005) have also proposed a method which helps speed up the classification stage by organizing the classifier into a binary tree to achieve a log($N$) time complexity. Tagare, Toyama, and Wang (2001) proposed a model in which an attentional strategy was used to reduce overall computations by performing fast but approximate image measurements. However, their computations involved finding parts of objects and determining their relationships in an approximate manner. In contrast, the contribution of the present paper is to provide a good feature set which can be quickly classified with a simple classifier, as well as the ability to use these feature sets to create a biased saliency map in order to quickly find the object in the scene regardless of pose. The methods described above can then be used to perform a more thorough evaluation of objects deemed by our system to be highly probable candidates, after these candidates have been selected in a first quick pass by our algorithm.

In this paper we draw inspiration from both the computer vision literature and models of the visual cortex and present a method based on a Bayesian framework to account for search and recognition in a probabilistic manner. In particular, a new model of combined attention and recognition is developed with dual emphasis. First, top-down biasing towards desired features should be readily available and, if possible, stronger than modulating the relative gains of different visual features guiding search, as explored in the past (Wolfe, 1994; Navalpakkam & Itti, 2007). Second, a common representational framework should be developed that can be used both for biasing towards desired targets as well as for speeding up recognition when these targets are found. We name our algorithm SalBayes which denotes our system's marriage of both saliency and Bayesian modeling.

From a biological aspect, this paper aims to develop a new approach which considers profiles of detectors that are more likely to respond to the target by shaping their tuning curve towards the target individually. In particular, we consider a Bayesian framework that uses the prior knowledge of the objects to help shape the response of the detector profile in a dynamic manner. This approach achieves greater granularity in the discrimination ability of the search without the added overhead and limitations of multiple sub-bands. Additionally, the same information learned during recognition is used to guide attention. This is achieved by learning the likelihood probability density functions (PDFs) of salient features of various objects and then using these likelihoods to compute a probable location of objects during a search task.

The result of this work is a single computationally efficient system which provides dual use. When given a location in an image, the system will output a sorted list of objects and the associated probabilities to the type of those object that can be found at the given location. Alternatively, when given a description of an object,

the system will produce a sorted list of locations and associated probabilities that the given object can be found at a particular location. From these results, other more comprehensive models (which would presumably be slower) can operate on these lists to yield robust object recognition and search. Hence, we address the problem of prioritizing search and recognition, narrowing down from long and unordered to shorter and ordered lists of candidates, rather than completely solving and outputting a single recognized object label at a single location. We show how this is achieved by learning the visual features of an object, which is used for recognition as well as for efficient top-down-guided search. In testing against large standard databases (Amsterdam Library of Object Images (ALOI) (Geusebroek, Burghouts, & Smeulders, 2005), Columbia Object Image Library (COIL) (Nene, Nayar, & Murase, 1996), and SOIL-47 (Burianek, Ahmadyfard, & Kittle, 2001)), we find that this approach delivers robust machine vision performance, comparable and much faster than other more sophisticated, computationally intensive, and state-of-the-art machine vision systems (HMAX, SIFT) for recognition, while additionally providing a common framework for search and recognition.

In the following section we describe the model and its components. We start with the simple problem of object classification, and of defining a representation that can be learned from example views of objects. We then explore how this representation can also be used to provide efficient visual search for the learned objects. Section 3 describes the testing methodologies, datasets and results, while Section 4 provides discussion of the model and results.

## 2. Methods

The model proposed in this paper draws its inspiration from Bayesian theory as well as from the bottom-up attention model proposed by Itti et al. (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000). By learning the statistical variations in features of various objects, the model is able to perform an efficient visual search for a given target object, as well as classify target and distractor objects. At its core, the model learns the probability of an object's visual appearance having a range of values within a particular feature map. In a visual search task, the model influences the various feature maps by computing the probability of a given target object for each detector within a feature map. As a result, locations in the maps with the highest probability will be searched first, as they indicate likely positions for the target object. Both the prior and likelihood probabilities can be learned from training views of the object and the context. As we will see, a chief advantage of this approach is in its simplicity and speed, which make it an ideal candidate for a front-end system that quickly narrows search down to a few likely candidates which can then be investigated in more detail by more sophisticated and time-consuming recognition algorithms.

### 2.1. Object representation

To uniquely describe the appearance of an object, a number of feature maps are computed from the biologically inspired bottom-up saliency model proposed by Itti et al. (Itti et al., 1998; Itti & Koch, 2000). The saliency map represents statistically unique locations in an image after being decomposed into different feature maps at several spatial scales. That is, the saliency map attempts to detect anomalies, or outliers in the image within various feature spaces. In this paper the feature map domains consist of intensity, color opponency (red–green, blue–yellow) and four orientations (0°, 45°, 90°, 135°). These particular feature maps were selected based on the implementation proposed by Itti et al. (1998) which derived its inspiration from a review of which elementary visual features contribute to visual saliency in natural scenes (Wolfe et al., 2004). In the absence of top-down modulations a normalization operator, $N(.)$, within each feature map weighs the values of detectors in a data-driven fashion based on their uniqueness in that map. That is, the more different the response of a given detector is from its neighbors and globally, the higher the weight assigned to that detector's output. This normalization operator can also be thought of as providing spatial competition between neighboring pixels. The normalizing operator is computed as follows (Itti et al., 1998):

1. Normalize the values in the map to a fixed range $[0\ldots,M]$, in order to eliminate modality-dependent amplitude differences;
2. Find the location of the maps global maximum $M$ and compute the average $\bar{m}$ of all its other local maxima; and
3. Globally multiply the map by $(M-\bar{m})^2$

The 42 feature maps (seven features at six spatial scales) are then combined into a saliency map, which indicates the saliency of each location in the image. Implementation details of this model have been described previously (Itti et al., 1998; Itti & Koch, 2000) and the algorithm is freely distributed in source code at http://iLab.usc.edu/toolkit/.

To characterize the target, the most salient features from each of the 42 feature maps are sampled within a given fovea size (or patch size) centered on the object. Note that this location does not need to be the center of the object, nor does the object need to be segmented. The only requirement is that the object should overlap with the fovea location. The spatial competition will help provide a consistent location from which to sample when the object undergoes various transformations (illumination direction, rotation, etc.). Selecting the most salient location to learn from also helps in searching for the object. For example, if we know that we are looking for a red dot on the object, then it's worth searching for a red dot.

The motivation behind sampling from the most salient location within each submap around the object is to select features that would uniquely describe the object, but would still remain invariant to transformations in illumination, rotation, translation, etc. This can then provide a very efficient search mechanism when attempting to narrow down possible objects during recognition. The argument follows that a salient location in an object would remain invariant to transformations since it is very unique to the object. For example, the model not only learns that the object has a particular strong color value, but also that it has a particular strong intensity, and particular orientations. Therefore, not only the conjunctions of various feature maps can yield a position that is highly salient, but also feature values within each feature map at these strong locations.

The method of only selecting particular key locations to describe objects and scenes, rather than considering the entire pixel array, has also been successfully used by the SIFT algorithm (Lowe, 2004) and has been studied by Mikolajczyk, Leibe, and Schiele (2005). However, this paper uses the saliency map described above which is a much more elaborate method of determining the keypoint locations in order to provide a more robust feature set for recognition. Note that only the single most salient location in each feature map is used to build the descriptor vector. This results in very quick recognition rates, since adding more locations would require a more complex model to account for spatial locations within them. In particular, the goal of the model is to code probable locations and or hypotheses of particular objects, but not determine them specifically. Therefore, we would want to use as few features with a few complexities in order to speed up initial recognition and search. Other, more complex models (which would require more time to compute), would then be used on these locations in order to specifically determine the object.

Nevertheless, since the features are sampled from multiple scales, some spatial information is encoded in the feature vector but is not tightly localized. This is a result of sampling from the various scale-space pyramids. Consequently, features extracted from a single salient corner of a rectangle will yield a different signature (vector of features) than a signature obtained from a square. This is due to the fact that a rectangle will occupy a different number of cells within the image, and thus will show up in a different pyramid level. Additionally, a more complex model (with multiple features and locations) can be considered but would result in less efficient recognition or search (especially when the spatial distributions of the features are included). For a similar approach to recognition with a more complex model (see Shokoufandeh, Marsic, & Dickinson, 1998).

During training, the object model descriptor is built by computing the likelihood probability distributions of the 42 features resulting from each feature map. This PDF is modeled using a Gaussian distribution for each individual feature type, where both the mean and variance are learned. That is, the algorithm learns 42 separate univariate Gaussian distribution for each object. The choice of this distribution is due to the simplicity and efficiency in obtaining the parameters mean ($\mu$) and variance ($\sigma^2$) in an on-line method from training images. Additionally, these likelihoods are used later for searching for the object. However, other distributions can be used such as super-Gaussians, mixtures of Gaussians, particle filters, or discrete histograms (Scott, 1992).

Given a region of interest patch $q$ with $N$ pixels from a particular location (this location will correspond to the image being trained with) from within a given feature map (from the 42 feature maps computed above), the spatial competition method $N(.)$ (the non-linear normalization method described above) is applied to this patch to form a new set of patch values $q'$. A feature vector $F$ is then built using the value of $q$ from the location at which $q'$ has a maximum response. This value then forms the $j$th component of the feature vector $F$, and is denoted $F_j$. In other words we select the center-surround feature that has the highest value in the spatial competition layer (the most unique feature in that map).

$$F_j = q[argmax(q'_i)_{i=1,\ldots,N}] \ \forall j \in F \tag{1}$$

where $i$ represents the pixel position within the patch, $F_j$ is the particular feature value from feature map $j$ and $F$ is the set of feature maps.

The Normal distribution is then used to estimate the likelihood, $p(F_j|\theta_j)$, of observing feature $F_j$ given a particular object class parameter for this feature $\theta_j$. For example, if $j$ is the index of the vertical Gabor detector channel, then $F_j$ would represent the response of that channel at it's most 'unique' location, as determined by $N(.)$. $\theta_j$ would then represent the learned mean and variance of the vertical Gabor responses for object $\theta$.

The final model ($\theta$) is then a set of $n$ parameters ($\theta_j$), each composed of a mean ($\mu$) and a variance ($\sigma^2$) for each individual feature map. This gives the ability to simply compute the model parameters ($\theta$) mean ($\mu$) and variance ($\sigma^2$) from the training views of the object within each feature map, and to use a Gaussian distribution to estimate the likelihood.

$$p(F_j|\theta_j) \propto N(F_j;\mu_j;\sigma_j) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{-(F_j-\mu_j)^2/2\sigma_j^2} \tag{2}$$

When learning from only a single view, the standard deviation ($\sigma$) is initially set to a fixed value of 0.001, which was chosen arbitrarily (this number should be small so that the particular feature detector will provide some discrimination). This gives the classifier a rough estimate of the classification of the object with only one training view of the object, while fully computing the variance requires more than one training view.

### 2.2. Object classification

To classify particular features obtained from the feature maps, a naive Bayesian network is used. The choice of a naive Bayesian network in the model was made to reduce the amount of computations necessary for classification, as this type of network assumes statistical independence between feature values. Since some of the features are derived from different scales in the image, our features are actually guaranteed to be statistically dependent. However, it has been shown that even if the features are statistically dependent upon one another, computing the full network often only increases accuracy by a small amount, whereas the computations necessary to achieve this small improvement are large (Rish, 2001). As further evidenced in Vasconcelos and Vasconcelos (2009), for image classification, modeling the joint distribution between pairwise features provides often only a marginal performance boost.

Once a set of features ($F$) is collected from a given salient location within the feature maps (as described above), the classification is performed using Bayes formula:

$$p(\theta_i|F) = \frac{p(F|\theta_i)p(\theta_i)}{p(F)} \tag{3}$$

To make a decision as to the type of classification assigned to an object, $i$ can be iterated over all known objects and the object with the greatest posterior can be chosen as the best match. This method is known as *Maximum a Posteriori* (MAP). However, the goal of the model is to act as a fast front-end to slower, more accurate object recognition systems, and so we instead output a list of objects and match likelihoods sorted by the probability that each object matches the requested location. In our experiments, the prior is taken to be $1/C$, where $C$ is the number of classes. This results in each class being equally probable to observe (uninformed prior). However, changing the prior in response to outside knowledge, could yield better classification rates if within a given scene the probability of a particular object appearing can be determined.

Since the probability of the evidence can be viewed as a normalizing constant (used to ensure that probabilities all add up to unity), it can be dropped from the equation. This is because the comparison of the posterior is between classes, and only the greatest one is selected and not its particular value (the scale of the value is insignificant). Furthermore, the assumption that features are statistically independent from one another simplifies the calculation to just multiplying the likelihoods together to come up with a decision, as opposed to calculating the full joint distribution between the features.

$$p(\theta_i|F) = \text{argmax}_i \left( p(\theta) \prod_{j=1}^{n} p(F_j|\theta_{ij}) \right) \tag{4}$$

Additionally, taking the product of many probabilities, some of which may be very small, can give rise to numerical instability. As a result, an underflow often occurs in a straightforward implementation of Eq. (3) when using more than a few features. A solution to this problem is to take the log of the likelihood which will convert the probabilities from being less than one to negative numbers greater than one. This also greatly simplifies the computations in our practical implementation, as likelihood products are transformed into likelihood summations. Also, the decision to select a suitable classification is not affected, since only the maximum of these values is considered. As a result of these various techniques, Eq. (3) can be described by the following formula:

$$p(\theta_i|F) = \text{argmax}_i \left( p(\theta) \sum_{j=1}^{n} \log(p(F_j|\theta_{ij})) \right) \tag{5}$$

The enhanced version of the saliency map with the Bayesian network used for object recognition can be seen visually in Fig. 3.

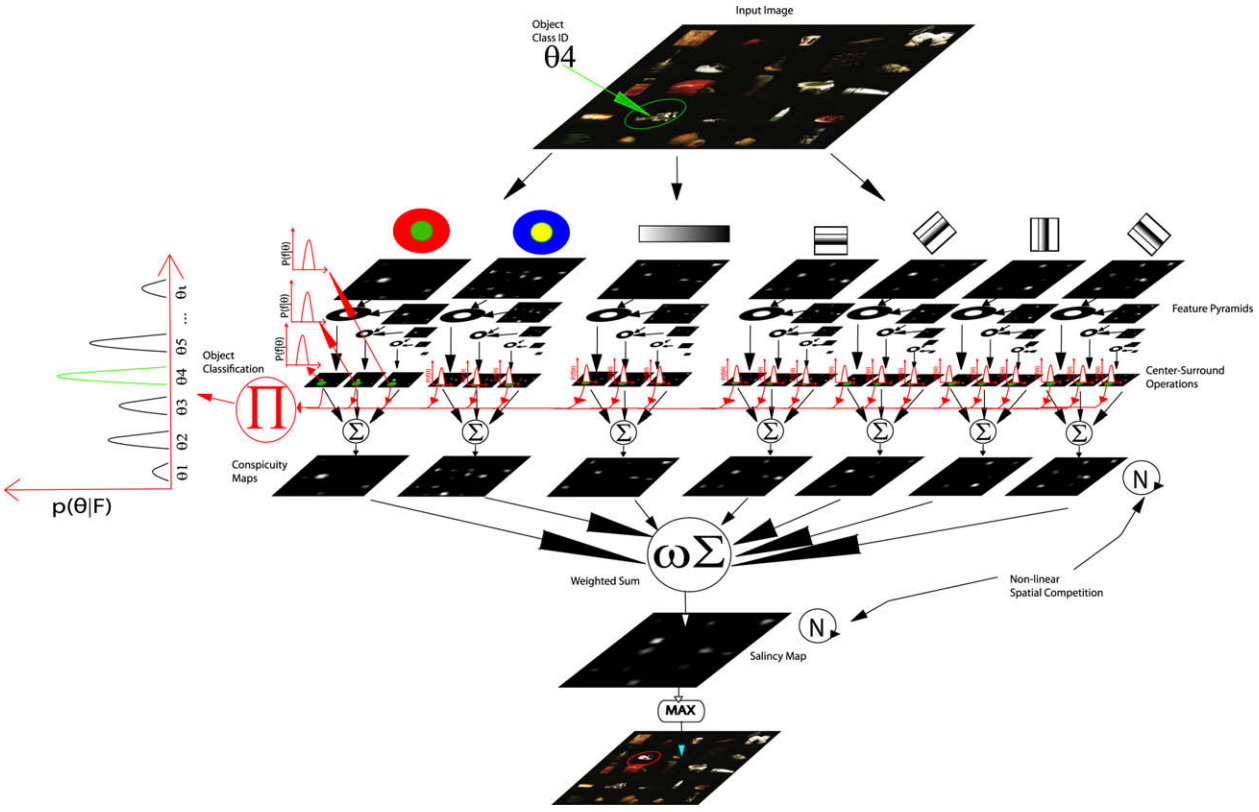



**Fig. 3.** The added Bayesian network to the saliency model for object recognition. Red indicates added components and data paths. The toy soldier at the input image is selected for learning/classification indicated by a green circle. The maximum feature location in each center-surround feature map is used to train or classify the Bayesian network for the selected object (indicated by the probability map on the left side). Each feature map builds a probability distribution of the most salient location in that map, shown in red. The rest of the architecture is as previously described.

### *2.3. Biasing learned features for efficient object search in a Bayesian framework*

Once the parameters of a particular object are known, they can be used to search for the object in an efficient manner. This is accomplished by biasing the feature maps to influence the saliency map so that the object that is being searched for becomes more salient, which can result in a faster search times simply by sorting by salience. For example, if our bottom-up saliency computations considered bright locations as salient, then darker locations would often be considered last as possible targets. However, if our object was dark in color, then biasing the saliency computations to choose darker locations as salient should improve search time, which would result in the biased saliency highlighting darker locations first.

The saliency map is biased by using the knowledge of the target parameters, and applying them to the set of feature detectors that are computed. Particularly, the parameters of our target are used to look for a particular mean and a variance within a given feature map. These parameters can be thought of as an envelope limiting the feature map response. In other words, the feature map would have its activation shaped by the likelihood of the particular feature value belonging to the object. Although our system could be thought of as only considering one sub-band, that sub-band can be dynamically shaped (regarding its position along the feature spectrum, and its specificity or width), thus providing an interesting alternative to using several fixed sub-bands. The result in the feature map then gives the probability of our object being coded by a particular feature detector. The maximum location within the feature map would then give an indication of the possible target location (Fig. 2c,d). The biasing process (applying the likelihood model to the feature map) is repeated within each feature map and the combination of all the feature maps' information is used to create a saliency map where the maximum indicates the most probable location of our target. The enhanced version of the saliency map with the Bayesian network used for finding objects can be seen in Fig. 4.

The various feature maps in the saliency map are biased in the following way: First the feature maps are computed by creating an image pyramid of each feature type and taking the difference between the pyramids to form center-surround responses at various scales as proposed in the original saliency algorithm (Itti et al., 1998; Itti & Koch, 2000). There are 42 such maps labeled $F_1 \cdots F_{42}$ (four orientations, one intensity, one blue–yellow, and one red–green all at six different scales). Note that spatial competition is not computed on these feature maps and just the raw center-surround values are used. However, it is important to remember that the spatial competition was used when extracting the feature values during the training phase. From the learned parameters of a particular object $\theta$ the parameters ($\mu_j$ and $\sigma_j$) corresponding to a particular feature map $F_j$ are used to calculate the probability of a particular detector belonging to the target $p(F_j|\theta_j)$ in feature map $j$. This is done across the $n$ different feature maps. The maps are then multiplied together (instead of the sum which was used in the original model) to yield the final saliency map. Therefore, the resulting saliency map calculates the probability of a location containing a given target ($p(F|\theta)$). Again, to avoid numerical instability and to speed up computation, the log of the probability is used.

$$\log(p(F|\theta)) = \log(\prod_{j=1}^{n} p(F_j;\theta_j)) \propto \sum_{j=1}^{n} log(N(F_j;\mu_j;\sigma_j)) \quad (6)$$

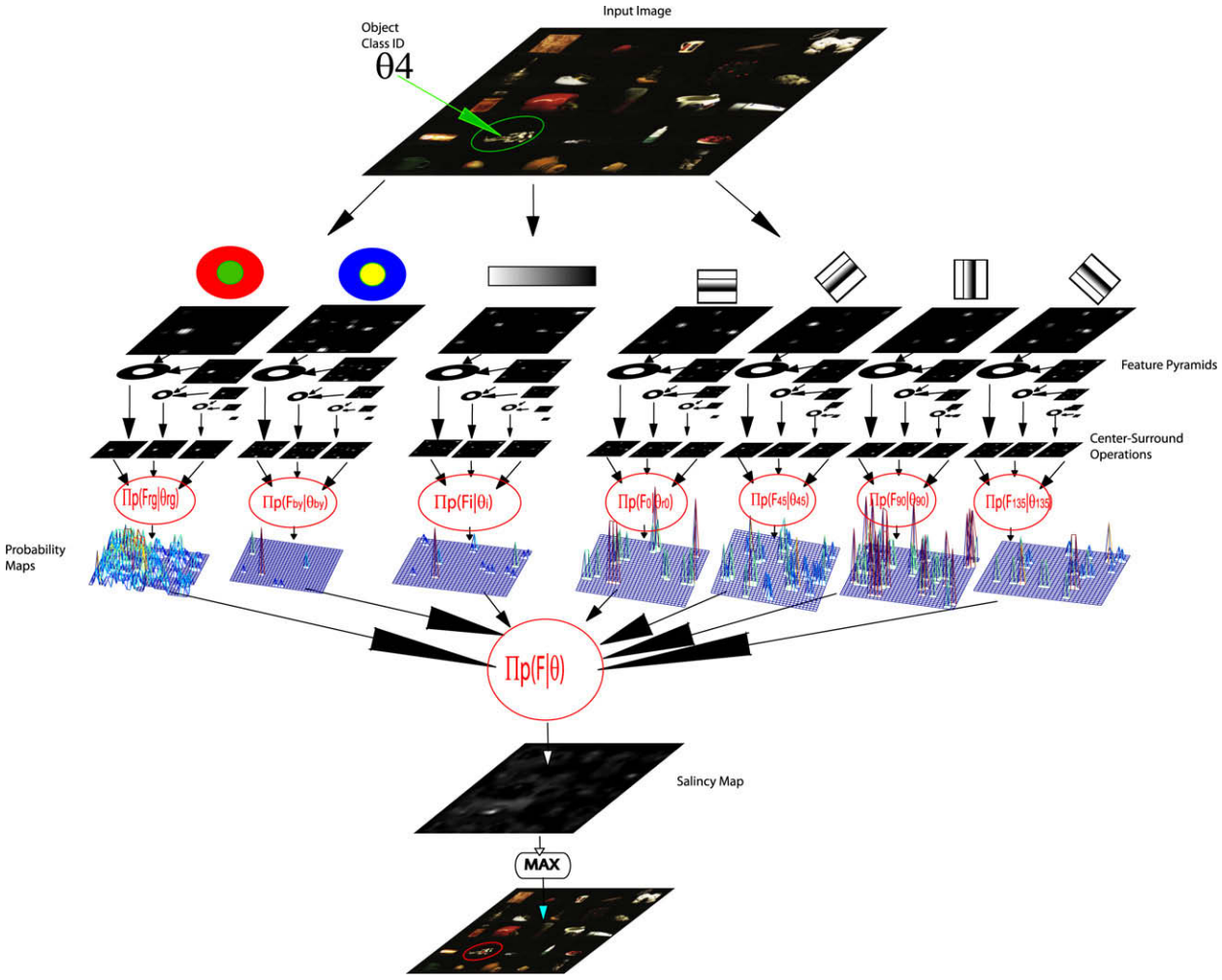



**Fig. 4.** The added Bayesian network to the saliency model for object recognition. Red indicates added components and data paths. The input image is passed for processing (without the selected object which is indicated in this image for clarity) by the saliency computations in the normal way. After the center-surround operations, the parameters of the object are used to find the probability of a detector indicating the position of the object in each submap (depicted as 3D graphs in the figure). All the submaps are then multiplied together to form the saliency map (note that in the implementation the multiplications are converted to additions by the used of the log operation).

$N$ is the normal function and $F$ is the set of all 42 features. Again, the spatial competition on the whole saliency map is not performed during object search. This is due to the nature of the spatial competition, which tends to punish high values within the same uniform region. Since that region describes the probability of the object, that location should be allowed to contribute to the overall saliency map.

Once a biased saliency map is computed, the locations with the highest locally maximal values in that map are searched first. That is, the object model is used on the locations which are local maxima in the biased saliency map. This processes in known as attention shifts. Finding local maxima is achieved by selecting the maximum value in the saliency map and applying an inhibition of return (IOR) mechanism to that location. The IOR is performed by applying a Gaussian disk mask with fixed radius to the saliency map which set all salient values underneath the mask toward zero, so that the next maximum salient location would have to be outside the disk. Implementation details of this mechanism have been described previously (Itti et al., 1998; Itti & Koch, 2000).

## 3. Results

The model was tested on three publicly available datasets to evaluate its performance in both object recognition and object search tasks.

### 3.1. Object recognition

For the object recognition task, several challenging datasets were used. These datasets included objects under many transformations including rotations and various viewpoints, illumination changes and illumination color changes. The original idea of the experiment was to use SIFT (Lowe, 2004) on top of the output of our model in order to speed up the search for object during recognition. However, during our preliminary experiments we found that using SIFT did not actually provide better recognition results than the raw output of our model. As a result, we directly compare the recognition capabilities of SalBayes against state-of-the-art object recognition methods: the SIFT (Lowe, 2004) algorithm as proposed by Lowe and the HMAX algorithm with feature learning proposed by Serre, Wolf, and Poggio (2005). These two methods were chosen due to their popularity in the machine vision and cognitive modeling community. For example, HMAX has been used to explain basic facts about the ventral visual system (Riesenhuber & Poggio, 1999) and has been used in object recognition (Serre et al., 2005, Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007), while SIFT has been used to build 3D models of objects (Snavely, Seitz, & Szeliski, 2006), robotics navigation (Se, Lowe, & Little, 2002; Elinas & Little, 2005; Sim, Elinas, & Griffin, 2005; Barfoot, 2005) and object classification (Lowe, 1999). Due to the large amount of data, a Beowulf cluster consisting of eight dual–dual-core Opterons(tm)

(four cores per node for a total of 32 cpus) running at 2.6 GHz was used to run the algorithms in a parallel fashion. The cumulative amount of CPU time taken for the testing sets was captured to compare the efficiency of the models.

The SIFT implementation was obtained from the author's website (Lowe, 2004), but the matching keypoints software was changed slightly to provide keypoint matching against a large dataset. In particular, a $k$-nearest neighbor algorithm (with $k = 2$) was used to determine the object identity given a test image and an image database. An implementation of HMAX with feature learning in Matlab was obtained directly from the author's website (http://cbcl.mit.edu/software-datasets/standardmodel/index.html). However, due to the large amount of data, the software was slightly modified to compute the features for all objects under all transformations and save them to a file first. This allowed us to extract the features in parallel using the Beowulf cluster. An SVM algorithm with a RBF kernel was used for training and testing. The implementation of the SVM was obtained from Chang and Lin (2001).

We test the proposed algorithm (along with HMAX and SIFT) against three large standard databases (ALOI, COIL, SOIL-47) separately and all together. The datasets are systematically broken into training and testing sets composed of the various images in the dataset. These sets include 1 image for training and the rest for testing, 6.25% training 93.75% testing, 12.5% training 87.5% testing, 25% training 75% testing and 50% training and 50% testing. The first object recognition dataset used was the Amsterdam Library of Object Images (ALOI) (Geusebroek et al., 2005). This dataset contains photographs of 1000 objects placed on a turntable and subjected to various transformations. These transformations include 12 illumination colors, 24 illumination directions, and 72 viewpoints (each object was rotated in steps of 5°). All photographs were first scaled down to a $256 \times 256$ pixel image to speed up computations. Several splits of the entire dataset into training and testing sets were considered, from using only one instance of each transformation (three images total) for training, to using half of the dataset for training. Object recognition testing was then performed on all 1000 objects on transformations that were not in the training dataset. The next data set used was the Columbia Object Image Library (COIL) (Nene et al., 1996) which consisted of photographs of 100 objects under 72 rotated views. The 7200 color images of $128 \times 128$ pixels were obtained by placing objects in the center of a turntable that was rotated at 5° increments. Here again several splits into training and testing sets were tested, from using only the first image for training and all others for testing, to using half of the dataset for training and the other half for testing. Object recognition was then performed on all 100 objects and on views that were not in the training datasets. The last dataset used was the SOIL-47 (Burianek et al., 2001) comprising photographs of 47 household objects. The images were obtained by placing a camera on a robot arm and moving it to various positions. In addition, the objects were also subjected to two illumination conditions. We again created training sets that ranged from just a single instance of each object, to half the dataset of the various views of the object. In addition, one of the illumination conditions for each of the two illumination conditions was used for training. Testing was then performed on all objects and on views that were not in the training datasets.

The results show that under many object transformations the model was able to successfully learn the objects, classify them correctly and search for them in an efficient manner. In particular, Fig. 5 and Table 1 shows that the model was able to classify the large datasets tested on average over 88.64% correctly. As indicated in Fig. 5, the HMAX algorithm was able to achieve a 92.46% classification rate on the ALOI dataset. Although this is a slight improvement over the proposed method, it should be noted that the features computed in the HMAX algorithm are 2000 dimensions in size and take more then 46 s to compute per image, as compared to the proposed model which uses only 42 features and is 279

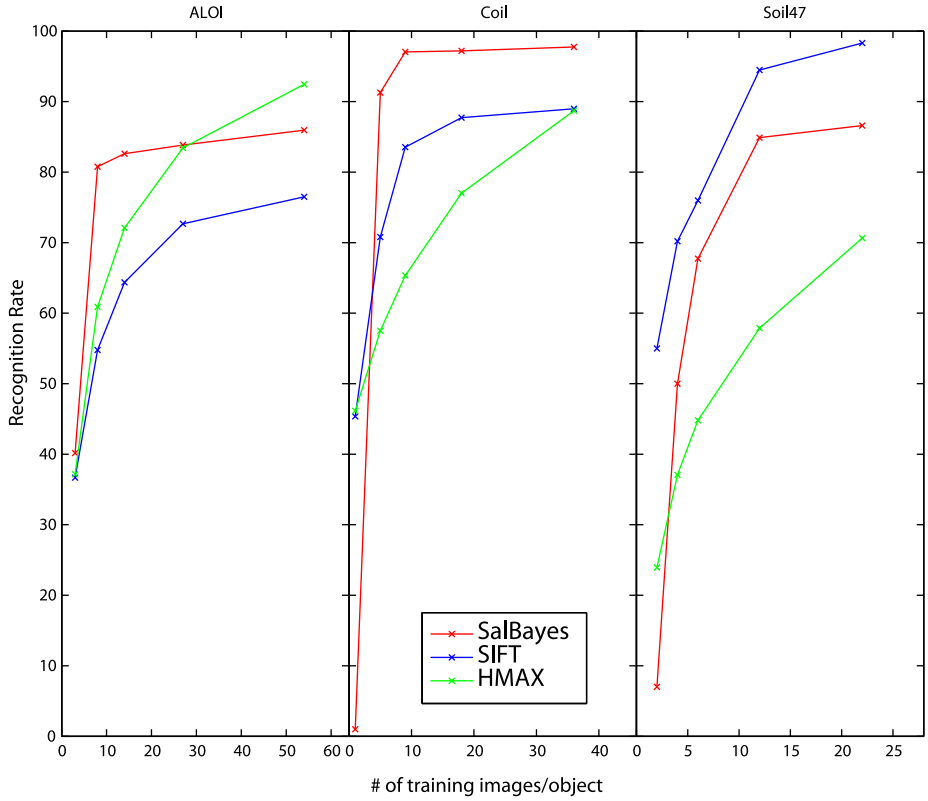



**Fig. 5.** Classification rates as a function of training size obtained by the proposed algorithm SalBayes, SIFT and HMax.

**Table 1**
Average recognition from the various datasets using 25% of the data for training. *N* represents the number of images in the testing set. The work of others have been included in this table to place the performance in context. To our knowledge, no one has before us used all of the 1000 objects in the ALOI database under all conditions.

| Method | Classification rate (%) | | | |
|---|---|---|---|---|
| | ALOI *N* = 81,000 | COIL *N* = 5400 | SOIL47 *N* = 1410 | ALL datasets *N* = 87,810 |
| SalBayes | 83.83 | 97.20 | 84.89 | 88.64 |
| SIFT | 72.68 | 87.19 | 94.48 | 84.78 |
| HMAX | 83.42 | 77.02 | 57.87 | 72.77 |
| MNS (Murthy, 2007) | – | 99.91 | 100.00[1] | – |
| LAF (Obdrzalek & Matas, 2002) | – | 99.90 | 100.00[1] | – |
| Graph matching (Kittler & Ahmadyfard, 2001) | – | – | 73.0 | – |
| Extra trees (Maree et al., 2005) | – | 99.50 | – | – |
| Sub-windows (Geurts et al., 2004) | – | 99.61 | – | – |
| SNoW/edge (Roth et al., 2002) | – | 94.13 | – | – |
| SNoW/intensity (Roth et al., 2002) | – | 92.31 | – | – |
| Linear SVM (Roth et al., 2002) | – | 91.30 | – | – |
| NN (Roth et al., 2002) | – | 87.50 | – | – |

**Table 2**
Recognition results on the ALOI dataset under the various transformations using 25% of the data for training. *N* represents the number of images in the testing set. The fifth column is the performance rate obtained when using all images (all images from A, B and C), while the sixth column represent an unweighted average of performance obtained for A, B and C (if the same number of transformations where equally likely to occur.).

| Method | Classification rate (%) | | | | |
|---|---|---|---|---|---|
| | A. Changes in illumination color only *N* = 9000 | B. Changes in illumination direction only *N* = 18,000 | C. Changes in rotation only *N* = 54,000 | All images from A, B, and C *N* = 81,000 | Unweighted average of performance for A, B and C |
| SalBayes | 64.79 | 75.50 | 89.71 | 83.83 | 76.67 |
| SIFT | 89.41 | 71.47 | 70.95 | 72.68 | 77.28 |
| HMAX | 99.04 | 83.13 | 80.76 | 83.42 | 87.64 |

times faster (0.165 s per image). Moreover, the increase in performance was only achieved when training on half of the dataset, which means that the difference between a training image and a testing image is not large. As seen, the proposed algorithm, SalBayes, was able to achieve better performance with less training data at speeds which greatly surpass both HMAX and SIFT. Examining the different datasets, it can be seen that the proposed model was able to learn the object features from only a few training examples (less then five per object) and achieve good results. In particular, the COIL dataset shows that from five training examples, the model was able to correctly identify 91.28% of the test images correctly. Lastly, the test result also show that the system performs fairly well when using only gray value images (just like HMAX and SIFT). This indicates that the proposed system can still provide useful information in the absence of color information.

Because the ALOI dataset contained the most systematic transformations, further analysis was done to determine the classification rate under each type of transformation. Looking at Table 2 we can see that HMAX performs best under several transformations. In particular, it does exceptionally well on the illumination color task. On the other hand, our new model performs well in the illumination color task when considering gray value images. Additionally, the model does exceptionally well under the rotation task. This shows the model's robustness against rotation and possible other 3D transformation (as can be seen in the soil47 dataset) as a result of picking the most salient features to remember when determining the classification of an object.

Looking at the timing aspects of the models tested, it can be seen that the proposed method, SalBayes, outperforms both SIFT and HMAX by many folds. Examining the results from Fig. 6, it can be seen that for testing on half of the ALOI dataset, it took only 3.42 h for the SalBayes algorithm as opposed to 4878.3 h for SIFT and 678.55 h for HMax. On average across all the datasets the new model was more than 1500 times faster than SIFT and 279 times faster than HMAX.

### 3.2. Grid based object search

The visual search task was evaluated by creating a dataset which consisted of search arrays created from the ALOI images.

Figs. 7 and 8 shows an example of a scene created for the search task. The scenes were created by taking random objects from the ALOI dataset under random transformations (from all 1000 objects) and placing them in a $2 \times 2$ or a $5 \times 5$ grid pattern. A random object was then chosen as the target object and the system searched for that target. This resulted in search images of size 512x512 pixels for the $2 \times 2$ grid and $1280 \times 1280$ for the $5 \times 5$ grid ($256 \times 256$ pixels per object). The parameters for the objects that were learned from training on half of the dataset as described above were used in this task. The number of "attention shifts" (inspections of individual objects) taken to reach the target object was then recorded. The inhibition of return (IOR) size was set to 30 pixels radius. This meant that only a small portion of the image would be inhibited at a time. As a result, multiple fixations per object could result if the object has strong multiple salient location, which would lead to greater number of fixations than the grid will allow.

Fig. 8 show the number of scenes vs. the number of attention shifts taken to reach the target object. The result show that during the search task, only 4.2 attended locations were required on average (with standard deviation of 5.9) to be examined in order to find the target object. About 218 fixations (290 fixations for 30 pixels IOR for the whole image minus 72 fixation for one 256 image) would be needed to systematically cover the whole image for the $2 \times 2$ and 1692 fixations for the $5 \times 5$ array. In particular, in these synthetically-generated scenes, the model was able to find the target object in fewer than five attended locations in over 76% of the scenes (average of $5 \times 5$ and $2 \times 2$ search arrays). Since the scenes could have contained any one of the 1000 objects, the ambiguity in the various scenes is large. For example, a few objects are green boxes, where the only varying feature is the size. Additionally, in

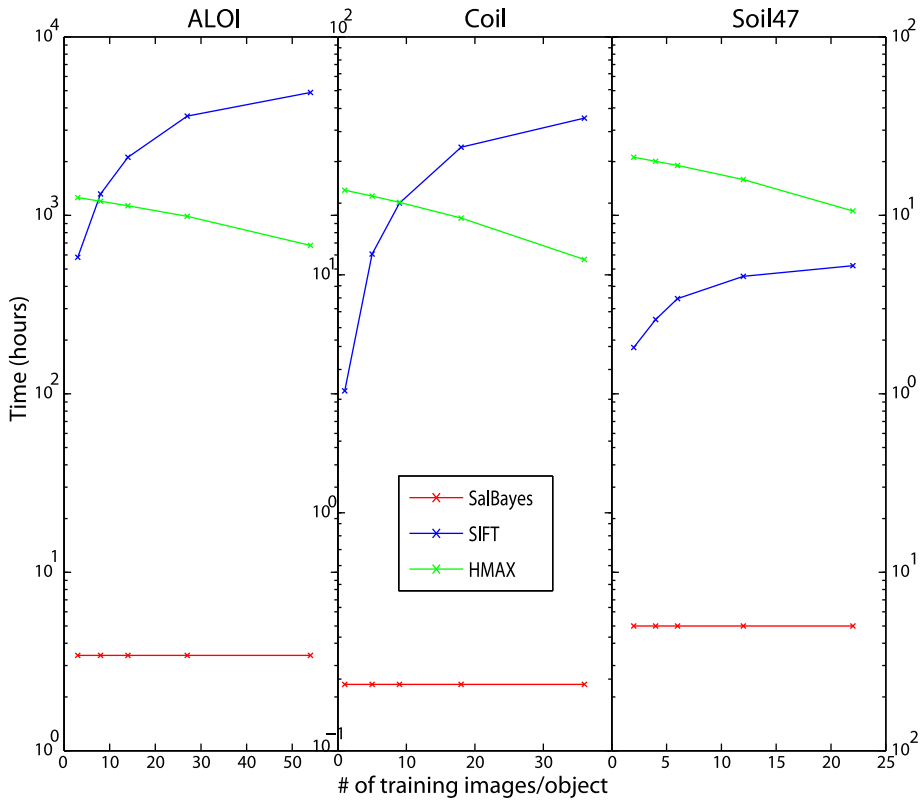



**Fig. 6.** Total CPU time required for testing, as a function of the fraction of each dataset that was used for training. As more images are used for training, fewer images remain for testing (hence the decrease in processing time for HMAX), but, in the case of SIFT, a larger keypoint database is built.

**Fig. 7.** Example 5 × 5 search scene built from the ALOI dataset. Scenes were created by taking random objects from the ALOI dataset under random transformations (from all 1000 objects) and placing them in either a 2 × 2 or a 5 × 5 grid pattern.

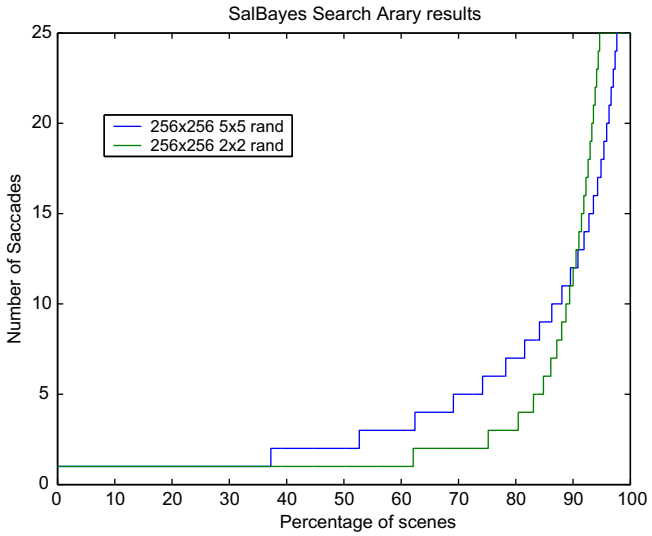



**Fig. 8.** Search results for the various scenes. The number of attention shifts (saccades) is plotted against the percentage of all scenes. For about 60 percent of the $2 \times 2$ scenes the object was located within the first fixation, and for about 38 percent of the $5 \times 5$ scenes the object was located within the first fixation.

some of the images, the object was never found due to a zero saliency value. Presumably, an exhaustive search would take place on the locations that were not searched.

### 3.3. Satellite image search

Another search task consisted of finding houses in satellite images. This task consisted of satellite images ($786 \times 786$ pixels) obtained from the New Orleans region after hurricane Katrina. An example application of this type of search would be to determine the number of houses effected by a natural disaster in an autonomous manner. This can be achieved by comparing the number of houses found before and after a disaster. Since satellite images contain a lot of data, it is often difficult for humans to quickly find places of interest in these images. In this task, the model was set to find images containing houses, so that humans can determine what do to with these regions (provide food, estimate the disaster area, etc.) The system was trained with 38 instances of houses obtained from 10 such satellite images ($786 \times 786$ pixels), and was tested on finding 95 houses spread out across 20 images. On an average each image contained five houses with standard deviation of 1.94 while occupying about $50 \times 50$ pixels. All images were hand labeled and a house was considered found if it was within a 30 pixel radius region of interest. For comparison, the same search task was used with the optimal gains proposed in Navalpakkam and Itti (2007).

Fig. 9 shows some of the training images used for the houses while Fig. 11 shows a typical satellite image upon which our model was used to find houses. To evaluate how well the Gaussian distributions fit the underlying probability distributions, the feature values were fit using a smoothing normal kernel function with a sliding window. The results shown in Fig. 10 indicate that the distributions do in fact resemble a Gaussian distribution. However, note that in some cases the distributions are highly peaked, which suggests that a super-Gaussian model may provide a slightly better fit.

As can be seen in Fig. 11, not all attended locations fell within houses, but the majority of locations did. On an average it took 1.52 searched locations with a standard deviation of 2.95 to find a house. The optimal gains method found a house within 1.95 searched location on average with a standard deviation of 1.51. Fig. 12 shows the percentage of the image that needed to be searched in order for find the houses in all 20 images. These results show that on average after searching about 25% of the image, all houses were found.

Fig. 12 also shows that the optimal gains method performed slightly better when finding the first few houses, but took much longer than our method to discover the more difficult targets. This slight improvement in initial performance is likely due to the fact that the optimal gains model considers both the target's and distractors' features in order to compute the best gain values. On the other hand, the SalBayes method only uses knowledge from the object to find the object. After finding a few houses, the performance of the optimal gains model drops considerably. This is mainly due to the max normalization method (see Section 2.1 and Navalpakkam & Itti, 2007 for details), which allows features which 'pop out' from the scene, yet are unrelated to the targets, to compete with those targets whose features are less visually unique.

## 4. Discussion

In this paper, we have developed a new unified model of attentional guidance and recognition which exploits the duality

**Fig. 9.** Training images used to find houses in the satellite images. The system was trained with 38 instances of houses obtained from 10 satellite images $786 \times 786$ pixels in size.

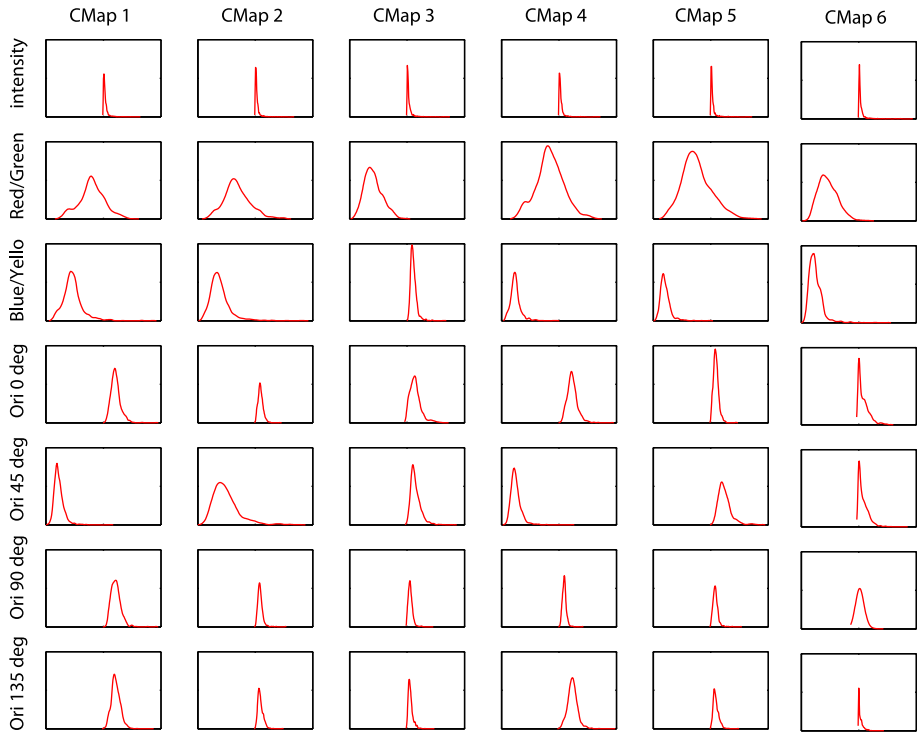



**Fig. 10.** Probability distribution of the houses for the various feature maps using a smoothing normal kernel function with a sliding window. The features are broken down in a grid where the rows indicate the feature type (intensity, color opponency (red–green, blue–yellow) and four orientations 0°, 45°, 90°, 135°), while the columns indicate the scale (1 being the coarsest and 6 being the finest).

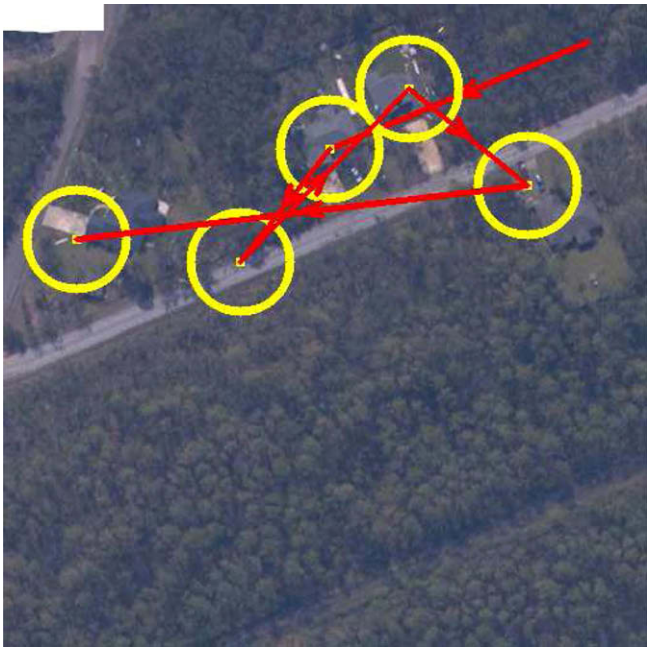


**Fig. 11.** Typical results for finding houses. The small yellow square indicates the fixation point, while the yellow circle indicates the inhibition of return size. The arrow shows the order in which the fixation points where chosen (which corresponded to saliency values). As can be seen, not all attended locations fell within houses, but the majority of locations did. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

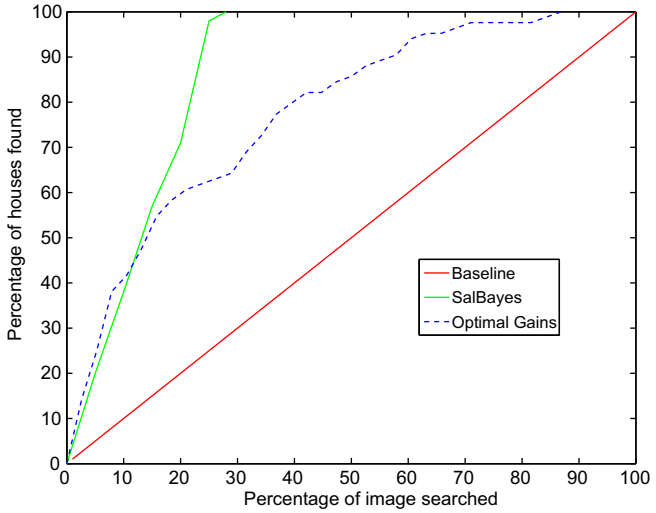



**Fig. 12.** Percentage of houses found vs. percentage of image searched for the 20 satellite images. Red line indicates the baseline performance if we tried to find houses at random. The green line indicates the performance achieved by SalBayes, while the dashed-blue line indicates Optimal Gains (see Navalpakkam and Itti, 2007). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between these two tasks. On the one hand, when the model is provided with a description of an object, it will output a probability map describing the likelihood that the object can be found at each location in an input image. On the other hand, when provided with only a location in an input image, the model will provide a list of probabilities denoting the likelihood that each of it's known objects is located at the given location. As shown in the results, the model performs informed search better than previous related efforts when given difficult targets, and has shown recognition performance that is on par with current state-of-the-art methods while providing very significant speed gains.

To our knowledge, no one has extensively tested their models across the three popular datasets used here all together. The work by Murthy (2007) comes close, but they only used a subset of the ALOI dataset for recognition. As can be seen, other models have been able to achieve superior performance on specific datasets. However, it is important to note that all of the successful methods use non-parametric methods for classification which causes their computation time to grow linearly in time with the number of training views. Since the goal of the proposed model is to provide a fast first layer recognition stage, any algorithm containing complex, non-parametric classifications will not be able to efficiently support a large object database. We propose that our model could easily lend it's vast speed improvements by operating as a fast front-end to such complex algorithms, and leave the analysis of such a hybrid system to future work.

Although the model was tested using an extensive dataset of objects and scenes, additional tests using objects in natural scenes could prove useful as well. However, there have not been any datasets created thus far which contain many objects (in the order of 1000) under systematic variations embedded in natural scenes.[1] Image databases such as the LabelMe (Russell, Torralba, Murphy, & Freeman, 2005) and Caltech 101/102 (Fei-Fei, Fergus, & Perona, 2006), do not provide a systematic object search, but are more concerned with general object search. For example, finding any chair could be viewed as search, but requires much more semantic knowledge for the search (there are many types of chair that could exist). As a result, a broad semantic meaning can cause great variations and ambiguity in the search. Additionally, most of the objects that people have labeled in the LabelMe dataset are salient to begin with and would not greatly benefit from a biased saliency map (Elazary & Itti, 2008). We believe that one of the strong points of our experimental validation in this paper is that it is very systematic, which will be more difficult to achieve with these type of labeled natural scenes.

Looking at the performance in the ALOI dataset against various transformations, we see that the model does not perform as well as HMAX under the illumination color condition. This is mostly due to the fact that the model considers color information to perform classification. Therefore, as the color of the object changes (due to the color of the light) the model encounters more ambiguity. However, such changes in color illumination do not often occur in the real world, and so we claim that robustness to 3D transformation and illumination direction are more desirable features in a first level recognition system.

Looking at the timing aspects of the methods tested, it can be seen that the proposed method, SalBayes, outperforms both SIFT and HMAX by many folds. Furthermore, the time requirement for both HMAX and SalBayes does not change significantly with training datasets (both decrease as the amount of remaining testing data decreases). This results from the underlying classifier that is used to classify the features. Both SalBayes and Hmax use a parametric density function to estimate the probability of the features belonging to a particular class. However, SIFT uses a non-parametric estimation ($k$-NN) which results in an increase in the time required to classify a given feature with the increase in training data.

While examining the performance of the proposed model, it was found that additional training examples did not always improve performance. This can be attributed to ambiguities developed by modeling each feature distribution as a unimodal Gaussian. When too many training instances are used, the actual distribution of a feature's density function can become multi-modal, which can then be poorly approximated by the model. Future work is planned to evaluate more advanced PDF representations, such as mixtures of Gaussians or particle filters to try to accommodate for such situations. However, despite these limitations, the proposed model has shown that from a very small dimensional feature vector (42 dimensions) at a single location on an object (the most salient location), the model was successfully able to distinguish among many objects.

One improvement to the model could be made by the choice of probability distribution. For example, after examining the features of particular objects, it was found that often the feature distribution could not be simply modeled using a single Gaussian model. That is, some of the variations of particular features could not be explained with a normal distribution. In particular the color (under the color illumination changes) and the orientations (under the rotation variations). As a result, estimating this as a normal distribution would cause errors in biasing and classifying the features. One explanation for the shape of these distributions can be due to the various ranges of values for different objects of the same class. For example, an object could contain strong red features and weak red features depending on the illumination color.

It was also found that the distributions in the ALOI dataset often exhibited two modes (which were primarily due to the changes in orientations and changes in illuminations). If the various variations of the objects can be modeled, then a single Gaussian can be used to describe a particular part of an object, and the mixture can be used to describe all the parts. Therefore, using a mixture of Gaussian model can provide a better model of the probability distribution. Training the mixture of Gaussian can be achieved by using an expectation–maximization (EM) algorithm. The drawbacks of this algorithm, however, are that it is an iterative method and requires that all training exemplars be available in each iteration. It would be worth investigating how the mixture of Gaussian model can be learned on-line as new inputs come in. One suggested way would be to cluster the data, extract the means, and then learn a single Gaussian on the cluster. The multiple clusters would then yield the mixture model.

Fig. 10 shows that some of the distributions in the satellite images house search could have been modeled using a super-Gaussian, to account for the sharp peak in the distribution. For example, the Laplace or logistic distribution could have been used in some of the distributions to model this peak. The results of which can improve performance by not only increasing the probability around the mean but accounting for more variations by having a fatter tail. However, future research will need to determine when and how to switch distribution models and how will this effect performance for both searching and recognition.

Examining the satellite images search results (Fig. 12), we see that the performance of the Optimal Gains proposed in Navalpakkam and Itti (2007) performs the same as the proposed model for the first few houses, but then loses performance when attempting to find more houses. The reason is that the Optimal Gains follows a similar structure proposed by Treisman's Feature Integration Theory (Treisman & Gelade, 1980; Treisman & Sato, 1990) and Wolfe's Guided Search (Wolfe, 1994) in which whole spatial maps of feature detectors are biased towards the target. Considering the neural hardware available in the brain (each neuron can perform computations independent of each other), it could be conceived that each neuron can be biased separately, which is the approach we have chosen to take in this paper. Additionally, we bias the feature maps with more of a probabilistic approach (applying a PDF for each neuron) as opposed to a simple gain change. This would enable the system to bias for weak features among strong ones as discussed in the introduction (since applying a gain would boost features and not suppress them). From a biological aspect, this can be seen as shaping the profiles of detectors that are more likely to respond to the target by shaping their tuning curve toward the tar-

[1] We are currently in the process of building an extensive dataset where the same objects are photographed in different complex natural backgrounds under different light conditions and poses.

get individually, using prior knowledge about the object. This results in great granularity in the discrimination ability of the search without the added overhead and limitations of multiple sub-bands.

Additionally, it is important to note that the optimal gains system as well as Feature Integration Theory and Guided Search is trained for search specifically, and does not uses this information for classifying the object. In this paper, we concentrated on the synergy between learning the parameters for the classification, and then using them for search. However, a hybrid system could be used so that the object can be searched more efficiently in the presence of known distractors. In particular, using some of the knowledge of the distractors could help achieve grater performance under certain conditions.

Lastly, the model proposed in this paper works in situations in which the object can be described using few simple features. For example, a house or a road can be described using simple features. However, more complex objects or scenes would need multiple features spanning greater spatial distance (more than the fovea size) to be described. For example, an urban area does not only contain a house but also contains multiple houses and roads (as seen from above). As a result, it would be advantageous if more knowledge can be added to the biasing, as proposed by Navalpakkam and Itti (2005). This knowledge would describe the parts of the object, and its relation in the scene. For example, if the system is looking for a refrigerator, then it knows that refrigerators are composed of doors. In addition, if the system is looking at a kitchen scene, then it can first check the likely locations of fridges within the scene first. Therefore, the knowledge of scenes can be used to efficiently speed up the search in more complex scenes. This knowledge can also boost the recognition rate by setting up the appropriate prior for the scene. For example, if we know the probability of a fridge appearing in a given scenes, ambiguities in appearances with another objects (say a door) can be resolved using the prior information about the scene. This knowledge can be provided from gist models, such as one proposed by Torralba, Oliva, Castelhano, and Henderson (2006). In addition, the knowledge base can be used to narrow down the search for features. For example, if a few houses are already encountered, then the system should check for the presence of trees. Therefore, the next fixation should bias for trees. As a result, the system knows that this could not be an ocean (because of the structure in the knowledge base), so it should not bias for boats. For a previous implementation of such a system (see Navalpakkam & Itti, 2005).

## Acknowledgments

This work was supported by the Army Research Office (ARO), the Human Frontier Science Program (HFSP), the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), and the National Geospatial-Intelligence Agency (NGA). The views, opinions, and findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency, National Geospatial-Intelligence Agency, or the U.S. Department of Defense.

## Appendix A

### A.1. HMAX

This visual structure was first proposed by Riesenhuber and Poggio (1999) and later improved by Serre et al. (2005). It was dubbed HMAX ("Hierarchical Model and X") and has drawn its inspiration from biological vision. The main contribution of the structure is its ability to achieve invariance at the local level by pooling local features using a max operator in both scale and position. The whole structure is built from two layers, where the first layer extracts Gabor features and pools them together. The pooling first takes the max over the position by sub-sampling the space into a grid size $N$-band and then taking the max between scales. The second layer extracts codewords at random from the first layer and stores them in a database. The response of the layer is then computed by a distance measure between the memorized patches and the current stimulus using Radial Basis function (RBF). Lastly, an SVM is used to classify objects based on the features from the second layer.

### A.2. SIFT

This algorithm has been proposed by Lowe (2004) and is known as SIFT, which stands for Scale-Invariant Feature Transform. The algorithm first extracts keypoints by using local scale-space maxima and minima of various Difference of Gaussian (DoG) operations applied to the input image. This results in keypoints from varius locations and scales with heigh texture energy. From these keypoints, a descriptor vector invariant to scale, translation, slight 3D rotations and intensity is created. This is achieved with a 128 dimensional vector indicating the gradient locations and orientations using a histogram. The space is quantizes into a $4 \times 4$ grid while the orientations are quantized into eight orientations. These descriptor vectors are store in a database for classification.

During the classification stage, the same processes described above is used to extract various descriptor vectors from a new image, while a Nearest Neighbor algorithm is used to find matches in the database. Additional, at least three close matching keypoints are required to match with an additional affine constraint (checked with an Hough transform) in order for the object to be recognized.

### A.3. SVM

Support vector machines (SVM) are a method of supervised classification and regression first proposed by Vladimir Vapnik in 1963 for linear separation. The hypothesis space of an SVM is a set of hyperplanes that attempts to achieve the largest distance to any sample in the training dataset for any class, which is known as the functional margin. To handle non-linear classification, SVMs employ a kernel trick proposed by Boser et al. (1992), which first maps the data into a liner space using a kernel of some kind, and then performs the linear separation. Common kernels include Polynomial, Radial Basis Function and Gaussian functions.

## References

Barfoot, T. (2005). Online visual motion estimation using fastslam with sift features. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), Edmonton, Alberta, 2–6 August 2005* (pp. 3076–3082).

Bonaiuto, J., Itti, L. (2005). Combining attention and recognition for rapid scene analysis. In *Proceedings of IEEE-CVPR workshop on attention and performance in computer vision (WAPCV'05), San Diego, California, June 2005* (pp. 1–6).

Boser, B. E., Guyon, I. M., Vapnik., V. N. (1992). A training algorithm for optimal margin classifiers.

Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal (Q2)*, 15. <http://citeseer.ist.psu.edu/bradski98computer.html>.

Burianek, J., Ahmadyfard, A., Kittle, J. (2001). Soil-47 the surrey object image library. <http://www.ee.surrey.ac.uk/Research/VSSP/demos/colour/soil47/>.

Chang, C.-C., Lin, C.-J. (2001). LIBSVM: A library for support vector machines. <http://www.csie.ntu.edu.tw/cjlin/libsvm>.

Comaniciu, D., Meer, P. (1997). Robust analysis of feature spaces: Color image segmentation. <http://citeseer.ist.psu.edu/comaniciu97robust.html>.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3:3), 1–15.

Elinas, P., & Little, J. J. (2005). sMCL: Monte-Carlo localization for mobile robots with stereo vision. In S. Thrun, G. S. Sukhatme, & S. Schaal (Eds.), *Robotics: Science and*

*systems I, June 8–11, 2005* (pp. 373–380). Massachusetts Institute of Technology, Cambridge, Massachusetts: The MIT Press. <http://www.roboticsproceedings.org/rss01/p49.html>.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 594–611.

Geurts, R. P., Piater, J., Wehenkel, L. (2004). A generic approach for image classification based on decision tree ensembles and local sub-windows. In: *Proceedings of 6th Asian conference on computer vision* (pp. 860–865).

Geusebroek, J. M., Burghouts, G. J., & Smeulders, A. W. M. (2005). The Amsterdam library of object images. *International Journal of Computer Vision, 61*(1), 103–112.

Gonzalez, R. C., & Wintz, P. (1987). *Digital image processing* (2nd ed.). Addison-Wesley.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194. <http://dx.doi.org/10.1016/j.tics.2005.02.009>.

Horn, B. K. P. (1986). *Robot vision*. Cambridge, MA: MIT Press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Kittler, J., & Ahmadyfard, A. (2001). On matching algorithms for the recognition of objects in cluttered background. In *Proceedings of the 4th international workshop on visual form (IWVF-4)* (pp. 51–66). London, UK: Springer-Verlag.

Krummenacher, J., Muller, H., Reimann, B., & Heller, D. (2001). Visual singleton detection ('pop-out') is mediated by dimension-based attention. In *Proceedings 17th annual meeting of the international society for psychophysics* (pp. 479–486). Pabst Science Publishers.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *ICCV* (pp. 1150–1157). <http://dx.doi.org/10.1109/ICCV.1999.790410>.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

MacLean, W. J., & Tsotsos, J. K. (2008). Fast pattern recognition using normalized grey-scale correlation in a pyramid image representation. *Machine Vision Applications, 19*(3), 163–179.

Maree, R., Geurts, P., Piater, J., & Wehenkel, L. (2005). Random subwindows for robust image classification. *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 34–40). Washington, DC, USA: IEEE Computer Society. <http://dx.doi.org/10.1109/CVPR.2005.287>.

Mikolajczyk, K., Leibe, B., & Schiele, B. (2005). Local features for object class recognition. *ICCV, 2*, 1792–1799.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782–784.

Motter, B. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area v4. *Journal of Neuroscience, 14*, 2178–2189.

Murthy, C. A. (2007). Distinct multicolored region descriptors for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(7), 1291–1296 (member – Sarif Kumar Naik).

Navalpakkam, V., Itti, L. (2006a). An integrated model of top-down and bottom-up attention for optimal object detection. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), New York, NY, June 2006* (pp. 2049–2056).

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205–231.

Navalpakkam, V., & Itti, L. (2006b). Top-down attention selection is fine-grained. *Journal of Vision, 6*(11), 1180–1193.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*(4), 605–617 (also see commentary/preview entitled "Paying Attention to Neurons with Discriminating Taste" by A. Pouget, D. Bavelier, Neuron 2007;53(4):473–475).

Nene, S. A., Nayar, S. K., Murase, H. (1996). Columbia object image library (coil-100).

Obdrzalek, S., Matas, J. (2002). Object recognition using local affine frames on distinguished regions. In Rosin, L. Paul, David Marshall (Eds.), *Proceedings of the British machine vision conference* (Vol. 1, pp. 113–122).

Obdrzalek, S., Matas, J. (2005). Sub-linear indexing for large scale object recognition. In *BMVC05* (pp. 113–122).

Pratt, W. (1991). *Digital image processing* (2nd ed.). New York: Wiley.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Rish, I. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*.

Roth, D., Yang, M., & Ahuja, N. (2002). Learning to recognize three-dimensional objects. *Neural Computation, 14*(5), 1071–1103.

Russell, B., Torralba, A., Murphy, K., Freeman, W. (2005). Labelme: A database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025, September 2005*.

Scott, D. W. (1992). *Multivariate density estimation*. Wiley.

Se, S., Lowe, D., Little, J. (2002). Global localization using distinctive visual features (October 2007).

Serre, T., Wolf, L., Poggio, T. (2005). Object recognition with features inspired by visual cortex. *Computer vision and pattern recognition (CVPR 2005), San Diego, USA, June 2005*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 411–426.

Shokoufandeh, A., Marsic, I., Dickinson, S. (1998). View-based object recognition using saliency maps. <citeseer.ist.psu.edu/48145.html>.

Sim, R., Elinas, P., Griffin, M. (2005). Vision-based slam using the raoblackwellised particle filter. In *Proceedings of IJCAI workshop on reasoning with uncertainty in robotics*.

Snavely, N., Seitz, S., Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3d. In *ACM transactions on graphics* (pp. 835–846). New York, NY, USA: ACM Press.

Tagare, H. D., Toyama, K., & Wang, J. G. (2001). A maximum-likelihood strategy for directing attention during visual search. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(5), 490–500.

Theeuwes, J. (1994). Endogenous and exogenous control of visual selection. *Perception, 23*, 429–440.

Theeuwes, J. (1995). Abrupt luminance change pops out; abrupt color change does not. *Perception & Psychophysics, 57*(5), 637–644.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766–786.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology – Human Perception and Performance, 16*(3), 459–478.

Treue, S., & Trujillo, J. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*, 575–579.

Tsotsos, J. K. (1991). Computational resources do constrain behavior. *Behavioral and Brain Sciences, 14*(3), 506.

Vasconcelos, M., & Vasconcelos, N. (2009). Natural image statistics and low-complexity feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(2), 228–244.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*, 202–238.

Wolfe, J. M. (1998). Visual memory: What do you know about what you saw? *Current Biology, 8*(9), R303–R304.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*(6), 495–501.

Image and Vision Computing 29 (2011) 1–14

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Visual attention guided bit allocation in video compression

Zhicheng Li [a,b], Shiyin Qin [a], Laurent Itti [b,*]

[a] School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[b] Computer Science Department, University of Southern California, Los Angeles, CA, USA

ARTICLE INFO

*Article history:*
Received 2 November 2009
Received in revised form 13 May 2010
Accepted 12 July 2010

*Keywords:*
Visual attention
Video compression
Eye-tracking
Video subjective quality

ABSTRACT

A visual attention-based bit allocation strategy for video compression is proposed. Saliency-based attention prediction is used to detect interesting regions in video. From the top salient locations from the computed saliency map, a guidance map is generated to guide the bit allocation strategy through a new constrained global optimization approach, which can be solved in a closed form and independently of video frame content. Fifty video sequences (300 frames each) and eye-tracking data from 14 subjects were collected to evaluate both the accuracy of the attention prediction model and the subjective quality of the encoded video. Results show that the area under the curve of the guidance map is $0.773 \pm 0.002$, significantly above chance (0.500). Using a new eye-tracking-weighted PSNR (EWPSNR) measure of subjective quality, more than 90% of the encoded video clips with the proposed method achieve better subjective quality compared to standard encoding with matched bit rate. The improvement in EWPSNR is up to over 2 dB and on average 0.79 dB.
© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

Significant improvements in video coding efficiency have been achieved with modern hybrid video coding methods such as H.264/AVC [1,2] in the last two decades. Spatial and temporal redundancy in video sequences has been dramatically decreased by introducing intensive spatial–temporal prediction, transform coding, and entropy coding. However, to achieve better compression performance, reducing such kind of so-called objective redundancy is limited and highly complex in computation.

On the other hand, research on human visual characteristics shows that people only perceive clearly a small region of 2–5° of visual angle. The human retina possesses a non-uniform spatial resolution of photoreceptors, with highest density on that part of the retina aligned with the visual axis (the fovea), and the resolution around the fovea decreases logarithmically with eccentricity [3]. What's more, research results show that observers' scanpaths are similar, and predictable to some extent [3]. These research results provide a new pathway to compress images/videos based on human visual characteristics: only encode a small number of well selected interesting regions (attention regions) with high priority to keep a high subjective quality, while treating less interesting regions with low priority to save bits.

Recently, many subjective quality-based video coding methods have been developed. According to the way of obtaining attention regions, they can be coarsely classified into four categories, as follows: (1) In the first approach, considering that human attention prediction is still an open problem, human–machine interaction methods are adopted to obtain the attention regions. One example of online human–machine interactive methods is gaze-contingent video transmission, which uses an eye-tracking device to record eye position from a human observer on the receiving end and applies in real-time a foveation filter to the video contents at the source [4–8]. This approach is particularly effective because observers usually do not notice any degradation of the received frames, since high-quality encoding continuously follows the high-acuity central region of the observers' foveas. However, this application is restricted to specific cases where an eye-tracking apparatus is available at the receiving end. For general-purpose video compression, this approach faces severe limitations if an eye-tracker is not available or several viewers may watch a video stream simultaneously. To address this, offline interactive methods are designed to obtain the interesting regions by asking subjects to manually draw regions which are interesting, and then applying this to the encoding procedure [9]. (2) The second class of approaches uses machine vision algorithms to automatically detect interesting regions. For instance, due to the importance of human faces while people perceive the world [10,11], it is reasonable to consider that human faces may likely constitute interesting regions. In [12–14], face regions are thus defined as the regions-of-interest. Face detection and tracking methods are explored to keep the interesting regions focused onto human faces, and more resources are allocated during encoding to these face regions, to keep these regions in high quality. With the development of face detection algorithms and object tracking methods in machine vision, this kind of video compression is very effective in the occasions where human faces indeed are central to the visual understanding of a video sequence, such as for video

* Corresponding author. University of Southern California - Hedco Neurosciences Building, room HNB-07A - 3641 Watt Way, Loa Angeles, CA 90089-2520 - USA. Tel./fax: +1 (213) 740 3527/5687.
*E-mail address:* itti@pollux.usc.edu (L. Itti).

0262-8856/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2010.07.001

telephone or video conference. However, this type of approach is obviously only workable when human faces are present. For unconstrained video compression where there may or may not be faces in the streams to be encoded, this method will fail to find interesting regions. (3) The third class of approaches uses knowledge about human psychophysics to guide the encoding process. For example, research results show that the human visual system (HVS) can tolerate certain amounts of noise (distortion) depending on its sensitivity to the source and type of noise for a given region in a given frame. Under certain conditions, the HVS can tolerate more distortion than the objective distortion measurements such as mean square error (MSE) would predict; on the other hand, there are some types of distortions which, despite low MSE, are vividly perceived and impair the viewing experience [15–17]. Based on this theory, many image/video encoding techniques have sought to optimize perceptual rather than objective (MSE) quality: these techniques allocate more bits to the image areas where human can easily see coding distortions, and allocate fewer bits to the areas where coding distortions are less noticeable. Experimental subjective quality assessment results show that visual artifacts can be reduced through this approach; however, there are two problems: one is that the mechanisms of human perceptual sensitivity are still not fully understood, especially as captured by computational models; the other is that perceptual sensitivity may not necessarily explain people's attention. For example, smoothly textured regions and objects with regular motions often belong to the background of a scene and do not necessarily catch people's attention, but these types of regions are highly perceptually sensitive if attended to. (4) The fourth class of approaches exploits recent computational neuroscience models to predict which regions in video streams are more likely to attract human attention and to be gazed at. With the development of brain and human vision science, progress has been made in understanding visual selective attention in a plausible biological way, and several computational attention models have been proposed [18–20]. In these models, low-level features such as orientation, intensity, motion, etc. are first extracted, and then through nonlinear biologically inspired combination of these features, an attention map (usually called saliency map) can be generated. In this map, the interesting locations are highlighted and the intensity value of the map represents the attention importance. Under the guidance of the attention map, resource can be allocated non-uniformly to improve the subjective quality or save the bandwidth [21–24]. Although such research shows promising results, it is still not a completely resolved problem.

Once interesting regions are extracted, a number of strategies have been proposed to modulate compression and encoding quality of interesting vs. uninteresting regions [21,25–29]. One straightforward approach is to reduce the information in the input frames. In [4,21,22], the frames to be encoded are first blurred (foveated) according to the attention map. The foveated image only keeps the attention regions in high quality while the other regions are all blurred. Through the blurring, redundancy is reduced significantly, and the compression ratio can be several times higher than the normal encoding method. However, blurring yields obvious degradation of subjective quality in the low saliency regions. In [23], a bit allocation scheme through tuning the quantization parameter is proposed with a constrained global optimization approach. Results show that 60% of the test video sequences encoded by this approach have better subjective visual quality compared to the video encoded by the normal method under the same bandwidth. In rate-distortion optimization, different mode may get different video quality and bit rate. The mode decision is usually determined by minimize the cost function which is the sum of encode error and bit rate multiple by a parameter (called Lagrange multiplier). Considering that the Lagrange multiplier will affect the mode decision in rate-distortion optimization, a Lagrange multiplier adjustment method is explored in [25]. An optimized rate control algorithm with foveated video is proposed in [26], and foveal peak signal-to-noise ratio (FPSNR) is introduced as subjective quality assessment. In [28], a region-of-interest based resource allocation method is proposed, in which the quantization parameter, mode decision, number of referenced frames, accuracy of motion vectors, and search range of motion estimation are adaptively adjusted at the macroblock (MB) level according to the relative importance (obtained from the attention map) of each MB.

How to evaluate the quality of a compressed image/video is still an open problem. Many quality assessment metrics have been developed to evaluate the objective or subjective quality of video. Among them, MSE and PSNR are two widely adopted objective quality measurements, even though they often are not consistent with human perception. Many additional types of objective (including human vision-based objective) quality assessment methods have been proposed [26,30–32]. However, the research results of the video quality experts group (VQEG) show that there is no objective measurement which can reflect the subjective quality in all conditions [33]. The suggested subjective quality from VQEG was obtained by using the mean opinion score (MOS) from pool of human subjects. Specifically, subjective quality scales ranging between excellent, good, fair, poor and bad (weight values are 5, 4, 3, 2, and 1, respectively) can be obtained from naive observers, and the weighted mean MOS score can be used as the subjective quality.

In this paper, we use a neurobiological model of visual attention, which automatically selects (predicts) high saliency regions in unconstrained input frames to generate a saliency map (SM). Considering the human's foveated retina characteristic, a guidance map (GM) is generated by finding the top salient locations in the saliency map. The GM is then used to guide the bit allocation in video coding through tuning the quantization parameters in a constrained optimization method. The overview of the proposed method can be seen in Fig. 1. For experimental validation, 50 high-definition ($1920\times1080$) video sequences were captured using a raw uncompressed video camera, which include scenes at a library, pool, road traffic, gardens, a dinner hall, lab rooms, etc. Instead of using a subjective rating method, an eye-tracking experiment which records human subjects' eye fixation positions over the video frames was conducted to validate both the attention prediction model and the compressed video subjective quality. The focus of this paper is to combine the attention model with the latest video compression framework, and to validate the result in a quantitative way through an eye-tracking approach. The experiment results show that the proposed method is effective in both predicting human attention regions and improving subjective video quality while keeping the same bit rate.

The present paper complements our previous work [21], in which we showed that a saliency map model can predict human gaze well above chance, and can be used to guide video compression through selective blurring of low-salience image regions. The key innovation in the present work is to replace the selective blurring step, which yields quite obvious distortions in low-salience video regions, with a more sophisticated and more subtle localized modulation of the H.264 encoding parameters. Our new algorithm employs a constrained global optimization approach to derive the encoding parameters at every location in every video frame. We find that the optimization can be solved in closed form, which gives rise to an efficient implementation. This new optimization approach is an important step as it yields encoded videos that subjectively look very natural and are not degraded by blurring. Further, we develop and test a new eye-tracking weighted PSNR (EWPSNR) measure of subjective quality. Using this measure, we find that videos compressed with the proposed technique have better EWPSNR on our test video clips. Because our proposed method is purely algorithmic, requires no human intervention or parameter tuning, is applicable to a wide variety of video scenes, and yields improved EWPSNR, we suggest that it could be integrated to future generations of general-purpose video codecs.

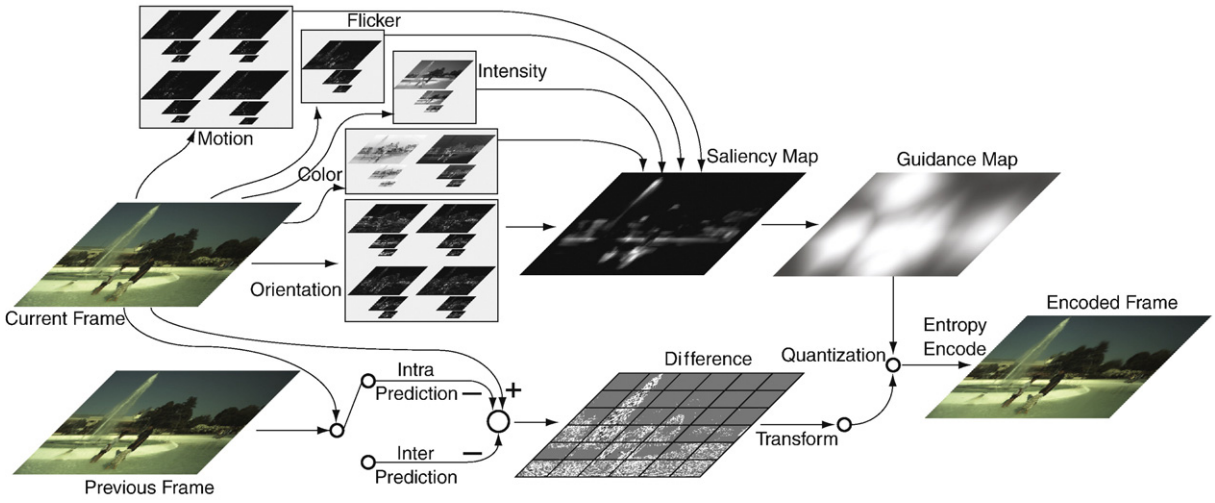



**Fig. 1.** Overview of the model. For the attention model path, the current input frame is first decomposed into multi-scale analysis with channels sensitive to low-level visual features (two color contrasts: blue–yellow, red–green; temporal flicker; intensity contrast; four orientations, 0°, 45°, 90°, and 135°; and for directional motion energies, up, right, down, and left). The saliency map is obtained after within-channel within-scales and cross-scales nonlinear competition. Assuming that the top salient locations in the saliency map are likely to attract attention and gaze of viewers, a guidance map is generated by foveating these positions. On the compression path, the current macroblocks (MBs) are predicted by previous encoded frame MBs through intra (which means the prediction result is generated from the current frame) or inter (which means the prediction result is generated from the previous frame) mode. The prediction error (difference) is then passed through transform and quantization; here the generated guidance map is used to adjust the quantization parameters to realize the non-uniform bit allocation. An encoded frame is complete after quantization and entropy encoding.

## 2. Method

### 2.1. Attention model

The model computes a topographic saliency map which indicates how conspicuous every location in the input image is. We used the freely available implementation of the Itti–Koch saliency model [34]. In this model, an image is analyzed along multiple low-level feature channels to give rise to multi-scale feature maps, which detect potentially interesting local spatial discontinuities using simulated center-surround neurons. Twelve feature channels are used to simulate the neural features which is sensitive to color contrasts (red/green and blue/yellow), temporal intensity flicker, intensity contrast, four orientations (0°, 45°, 90°, and 135°) and four oriented motion energies (up, down, left, and right). The particular low-level features extracted here have been shown to attract attention in humans and monkeys, as had been previously investigated in details [19,35,36]. Center-surround scales are obtained from dyadic pyramids with 9 scales, from scale 0 (the original image) to scale 8 (the image reduced by factor to $2^8 = 256$ in both the horizontal and vertical dimensions). Six center-surround difference maps are then computed as point-to-point difference across pyramid scales, for combination of three center scales ($c = \{2,3,4\}$) and two center-surround scale differences ($\delta = \{3,4\}$). Thus, six feature maps are computed for each of the 12 features, yielding a total of 72 feature maps. Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition [37]. In this way, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. All feature maps finally contribute to the unique scalar saliency map, which represents visual conspicuity of each location in the visual field.

After the saliency map is computed, a small number of discrete virtual foveas endowed with mass/spring/friction dynamics attempt to track a collection of most salient objects, using proximity as well as feature similarity to establish association between $n$ salient locations and $p$ fovea centers (similar to the approach described in our previous work [21,22]). The association is established through an exhaustive scoring of all $n \times p$ possible parings between a new salient location $X_t^A(i) = (x_t^A(i),\ y_t^A(i))$, $i \in \{1...n\}$ and an old foveation center $X_t(j) = (x_t(j),\ y_t(j))$, $j \in \{1...p\}$ at time $t$. (Typically, $p$ is fixed and $n = p + 4$ to ensure robustness against varying saliency ordering from frame to frame, $p = 10$ in the present implementation). Four criteria are included to determine the correspondence: (1) Euclidean spatial distance between the locations of $i$ and $j$; (2) Euclidean distance between feature vectors extracted at the locations of $i$ and $j$ which coarsely capture the visual appearance of each of the two locations; (3) a penalty term $|i-j|$ that discourages permuting previous pairings by encouraging a fixed ordered pairing; and (4) a tracking priority that increases with salience, enforcing strong tracking of only very salient objects. Combining these criteria tends to assign the most salient object to the first fovea, the second most salient object to the second fovea, etc. Video compression priority at every location is then related to the distance to the closest fovea center (using a 2D chamfer distance transform). For implementation details, please refer to [21] and [34]. It is important to note that the dynamics of the virtual foveas do not attempt to emulate human saccadic eye movement but track salient objects in a smooth and damped manner. The adopted correspondence and tracking algorithm compromises between reliably tracking the few most salient objects, and time-sharing remaining foveas among a larger set of less salient objects.

### 2.2. Bit allocation strategy

Assume that the rate-distortion ($R$-$D$) function is as follows [23,38] for a given region (typically, macroblock) $i$ in an image:

$$D_i(R_i) = \sigma_i^2 * e^{-\gamma R_i} \tag{1}$$

in which $D_i$ denotes the mean square error, $R_i$ stands for the bit rate, and $\sigma_i^2$ is a measurement of the variance of the encoding signal (both spatial and temporal) and describes the complexity of the video content, $\gamma$ is a constant coefficient. This approach assumes that the distortion is to be computed in a uniform manner, i.e., distortion in different areas of an image is equally important. However, if we take

the human's visual spatial resolution into consideration, then the encoding distortion in area $i$ can be written as follows:

$$D_i^{'} = w_i * D_i \tag{2}$$

here $w_i$ is the weight coefficient, it stands for the human's spatial resolution in area $i$. In the area around the center of gaze (the fovea), $w_i$ should be higher than in areas far from the gaze position, because even small distortions in the foveal region can cause people's awareness while in peripheral regions relatively larger distortions may not catch people's attention.

According to this non-uniform distribution of human eye resolution, there are two ways to optimize the bit allocation in encoding. One is as proposed in [23]: keep the sum of bit rate as a constant value and maximize the subjective quality. Under the hypothesis that $\sigma_i^2$ are equal at every location, the conclusion is then that the quantization parameter $QP_i$ is inverse exponentially with the optimized bit rate:

$$R_i = R + \frac{1}{\gamma NS}\sum_{j\neq i} S_j \cdot \log\frac{w_i}{w_j} (i = 1, 2...N) \tag{3}$$

$$QP_i = e^{\frac{\alpha - R_i}{\beta}} \tag{4}$$

where $S$ is the area size of the entire frame, $S_i$ is the area of region $i$, $N$ is the region number in one frame. $\alpha$ accounts for overhead bits and $\beta$ is the adjustment parameter. However, in reality, $\sigma_i^2$ are quite different over space within one frame and it is hard to determine the parameters correctly. In [23], the parameters are calculated from training videos that are similar to those to be encoded by the system, which is time consuming and may be unreliable if the test set differs substantially from the training set. To avoid this, we take a different approach: preserve the subjective quality while minimizing the bit rate. With this method, we find that the optimized distortion distribution is independent of the video frame contents and only depends on the weight coefficients. The details of this new method are described as follows:

To minimize the bit rate while keeping the subjective quality the same, we can write this global optimization problem as follows:

$$\begin{cases} Min \sum_i S_i * R_i / S \\ s.t. \sum_i w_i * D_i / W = D \end{cases} \tag{5}$$

here $W$ is the sum of all of the weight coefficients in different areas, $D$ is the target distortion. With the Lagrange multiplier method we can solve this equation in closed form:

$$\begin{cases} f(D_1, D_2, ...D_N) = \sum_i s_i * R_i / S + \lambda(\sum_i w_i * D_i / W - D) \\ R_i = \frac{1}{\gamma}(\log \sigma_i^2 - \log D_i) \end{cases} \tag{6}$$

in which $N$ is the number of areas in the encoded image. To obtain the minimum value, we pose that, at the minimum:

$$\frac{\partial f}{\partial D_1} = \frac{\partial f}{\partial D_2} = ... = \frac{\partial f}{\partial D_N} = 0 \tag{7}$$

Solving these equations above, we obtain:

$$D_i = \frac{W * s_i}{w_i * S} * D \ (i = 1, 2, ..., N) \tag{8}$$

We can see from this equation that the optimum $D_i$ is independent of the video characteristic related parameters $\gamma$ and $\sigma_i$. Hence the optimum process can be applied to any video no matter what the content of the video is, and we need not train on any sample videos to compute the optimum parameters. Furthermore, from the Eq. (8) we can see that the optimized distortion should be inversely proportional to the weight coefficient. One special condition is when all weights are equal over all locations, in which case the distortion should be equally distributed.

Now we can determine the bit allocation strategy with the calculated distortion distribution. In mainstream video compression schemes developed so far, the distortion stems only from quantization. The basic quantization operation is as follows:

$$Y = \text{round}\left(X / Q_{step}\right) \tag{9}$$

where $X$ usually is the coefficient after the transform (DCT, DWT, etc.), $Q_{step}$ is the quantization step and $Y$ is the quantized result. The $Q_{step}$-distortion mapping is linear and we can simply write the $Q_{step}$-distortion (Q-D) model as follows:

$$D = k * Q_{step} \tag{10}$$

where $k$ is a constant coefficient related to the video content. Then the optimized distortion for each area transformed to the optimized quantization step is, at every location $i$:

$$Q_{istep} = \frac{W * s_i}{w_i * S} Q_{step} \tag{11}$$

The formula above shows that in the human visual characteristic based video coding, the quantization step should be inversely proportional to the subjective quality weight coefficient.

According to the analysis above, we can apply the GM to guide the quantization parameter adjustment to conduct the optimized bit allocation. The GM values can be seen as the subjective weight coefficients and the quantization parameters then can be computed from above formula.

## 3. Video acquisition

Fifty video clips ($1920 \times 1080$) were collected for this experiment and each of them was cut to 300 frames (Fig. 2). All these clips were captured by a Silicon Imaging SI-1920HD camera with an EPIX E4 frame grabber card at frame rate of $30 \pm 0.2$ fps. The original frames were captured as Bayer format without any compression and saved in round-robin onto 4 separate hard drives to avoid limitations in frame rate due to limited disk bandwidth. The clips were captured around the USC campus and include both outdoor and indoor scenes at daytime. The outdoor scenes include library, pool, traffic road, gardens, museum, park, gates, fountains, square, lawn, track & field, and the indoor scenes include dinner hall, lab rooms, etc. After video capture, all the frames were converted to RGB color images through linear interpolation [39] and enhanced by gamma correction. Finally the frames were assembled into video clips in the YUV (4:2:0) format for further processing.

To facilitate future research, this raw uncompressed video dataset, as well as all the eye-tracking data described below, are made freely available on the Internet (http://iLab.usc.edu/vagba/). We hope that this comprehensive dataset and the associated human eye movements can benefit a number of research projects aiming at improving video compression quality, and beyond.

## 4. Human eye-tracking

The collected 50 uncompressed YUV format video clips were presented to 14 subjects and their eye fixation points were recorded over frames from each clip by an eye-tracker machine. The recorded eye traces represent the subjects' shifting overt attention, thus the

**Fig. 2.** Example of captured frames, first row: *dance01*, *seagull01*, and *garden09*, second row: *road02*, *fountain01* and *robarm01*, third row: *park01*, *gate03* and *lot01*, fourth row: *foutain05*, *room02* and *field06*.

eye-tracking data are qualified to validate the performance of the attention prediction model and the visual subjective quality.

Subjects were naïve to the purpose of the experiment and had never seen these video clips before. They were also naïve to attention theory, saliency theory, and video compression theory. They were USC students and staff (7 males, 7 females, mixed ethnicities, ages 22–32, normal or corrected-to-normal vision). They were instructed to watch the video clips without any specific task, and to attempt to follow whatever interesting things they might like. Later they were asked some general questions about what they had watched. The motivation of these instructions was to try to make the experiment similar to ordinary video watching. We believe that our instructions did not bias subjects toward low-level salient frame locations as defined by the Itti–Koch Model [18].

Stimuli were presented on a Sony Bravia XBR-III 46″ 60 Hz 1080 p LCD-TV display connected to a Linux computer. Subjects were seated

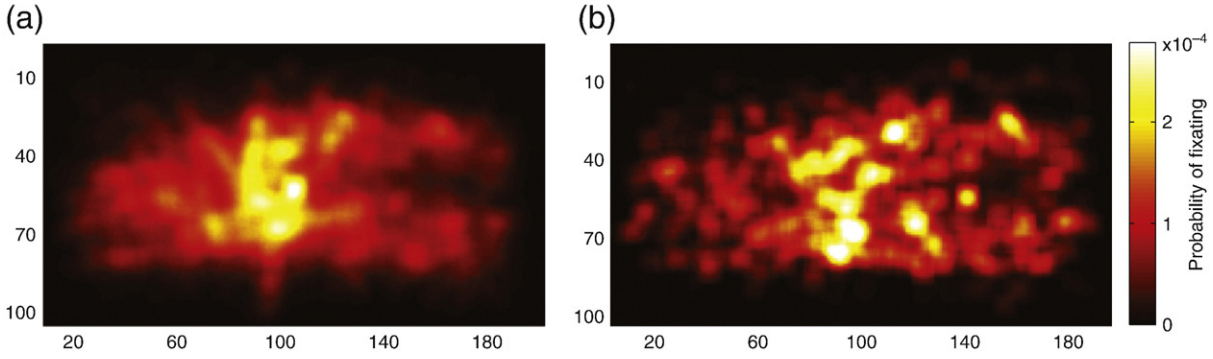



**Fig. 3.** Eye fixation distribution examples. The maps are histogrammed over $10 \times 10$ image tiles and normalized to 1. (a) The overall distribution for all subjects and clips shows a bias toward the center of the display. (b) Eye fixation distribution from one of the subjects over all the clips.

on an adjustable chair at a viewing distance of 97.8 cm, which responded to a field of view of 54.8 × 32.7°, and rested on a chin-rest. A nine-point eye-tracker calibration was performed every ten clips. Each calibration point consisted of fixating first a central cross, then a blinking dot at a random point on a 3 × 3 matrix covering the screen area. For each clip, subjects first fixated a central cross, pressed a key to start, at which point the eye-tracker was triggered, the cross blinked for 1066 ms, and the clip started. After each clip, the display became grey and the eye-tracker was disabled. The experiment was self-paced: the next clip was shown when the subject pressed the space button. Every ten clips, subjects could stretch before the nine-point calibration. Stimuli were presented on a Linux computer, under SCHED_FIFO scheduling (process would keep 100% of the CPU as long as needed) to guarantee timing. Each uncompressed clip (1920 × 1080, YUV 4:2:0 format) was entirely preloaded into memory. Frame displays were hardware-locked to the vertical retrace of the monitor (one movie frame was shown for two screen retraces, yielding a playback rate of 30.00 fps). Microsecond-accurate time-stamps were stored in memory as each frame was presented, and later saved to disk to check for dropped frames. No frame drop ever occurred and all timestamps were spaced by 33.333 ± 0.001 ms. Eye position was tracked using a 240-Hz infrared-video-based eye-tracker (ISCAN, Inc., model RK-464). Methods were similar to previously described [21]. In brief, this machine estimates point of regard (POR) in real-time from comparative tracking of both the center of the pupil and the specular reflection of the infrared light source on the cornea. This technique renders POR measurements immune to small head translations (tested up to 10 mm in our laboratory). All analysis was performed offline. Linearity of the machine's POR-to-stimulus coordinate mapping was excellent, as previously tested using a 7 × 5 calibration matrix in our laboratory, justifying a 3 × 3 here. The eye-tracker calibration traces were filtered for blinks and segmented into two fixation periods (the central cross, then the flashing point), or discarded if that segmentation failed a number of quality control criteria. An affine POR-to-stimulus transform was computed in the least-square sense, outlier calibration points were eliminated, and the affine transform was recomputed. If fewer than six points remained after outlier elimination, recordings were discarded until the next calibration. Otherwise, a thin-plate-spline nonlinear warping was then applied to account for any small residual nonlinearity. Data was discarded until the next calibration if residual errors greater than 34 pixels (about 1° field of view) on any calibration point or 17 pixels (about 0.5° field of view) overall remained. Eye traces for the ten clips following a calibration were remapped to screen coordinates, or discarded if they failed some quality control criteria (excessive eye-blinks, motion, eye wetting, or squinting). Calibrated eye traces were visually inspected when superimposed with the clips.

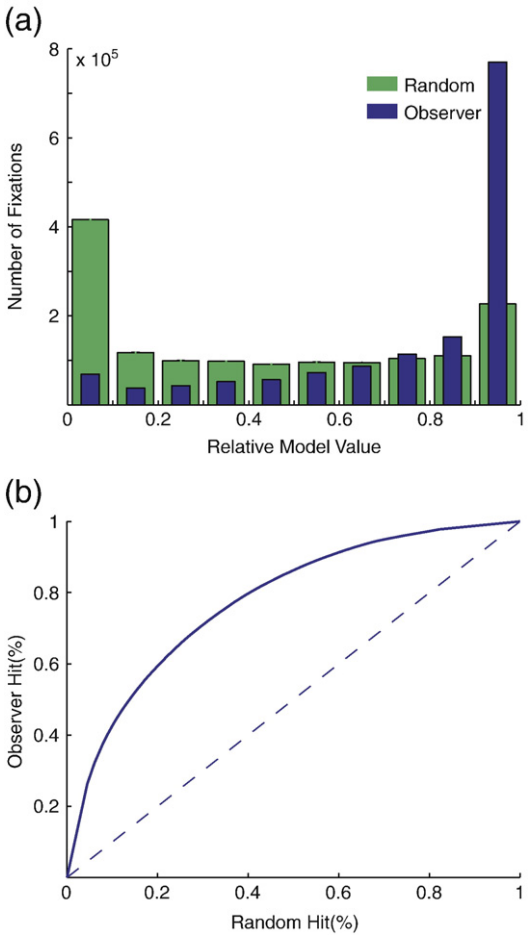



**Fig. 5.** Ordinal dominance analysis, there are 1,455,279 fixation points in total. (a) Histogram of guidance map values at eye positions and random locations. (b) Ordinal dominance curve, the dashed line is the chance level.

Eye fixation distribution examples can be seen in Figs. 3 and 4. We can see from Fig. 3 that the overall distribution of eye fixations is quite strongly center-biased on average, however, for different content clips, the eye fixation distributions are totally different and not necessarily center-biased (Fig. 4). This is important as it suggests that a simplistic saliency map – which would simply mark central screen regions as more salient – may work well on average but not necessarily for individual video clips (see [21] for

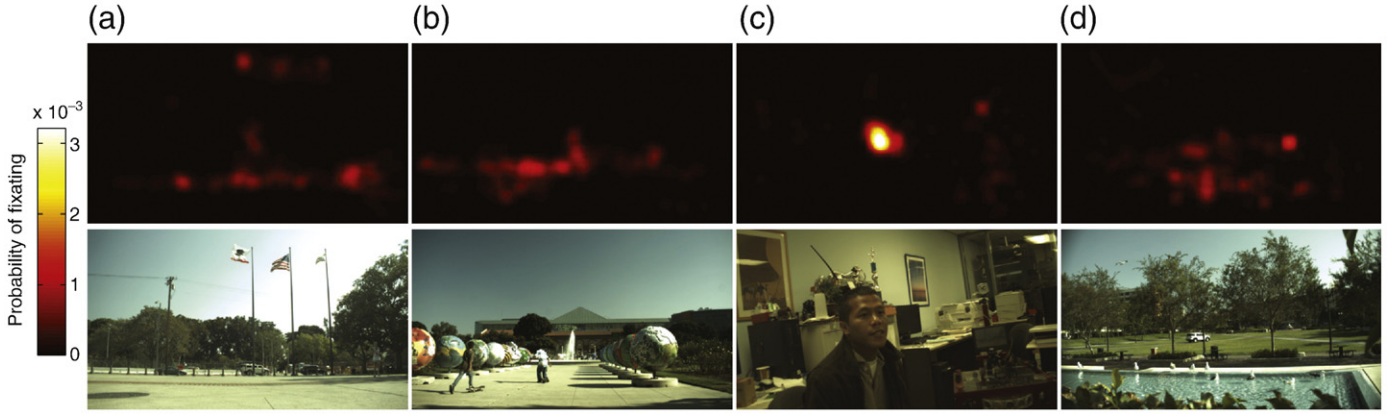



**Fig. 4.** Examples of eye fixation distribution map at different clips, the maps are histogrammed over 16 × 16 image tiles and normalized to 1. (a) *gate03*, (b) *park01*, (c) *room02*, (d) *seagull01*.

further discussion). In our results below we not only report average performance but also performance on individual clips and worst-case performance.

## 5. Experiment result

Uncompressed videos were shown to the subjects and the eye-tracking data was used to validate both the attention prediction model and the attention-based bit allocation scheme.

Considering that a good attention prediction model should output a model map which highlights the eye's fixation point, differences between guidance map values at subjects' gaze targets and at randomly selected locations were quantified, to evaluate performance, using ordinal dominance analysis [40]. Model map values at subjects' fixation points and at randomly selected locations were first normalized by the maximum value in the map when the eye fixation occurred (100 random locations are selected in this paper). Then, histograms of values at eye positions and random locations were created. Fig. 5(a) shows the results of the histograms over all video clip types and subjects. It is easy to see from the figure that many more human fixations were to high model salience values than expected by chance, which means the proposed model performs better than the random model at predicting human gaze. The mean observer value at the guidance map is 0.778 while the median value is 0.922, compared to mean 0.435 and median 0.400 obtained with random fixations (both model values are significantly higher than random value, $p<10^{-20}$, considering both $t$-tests for mean value comparisons and sign tests for median value comparisons). Fig. 6 shows the example of frames and their corresponding guidance maps. The subjects' eye fixation points are marked as small color patch in the frames.

To further measure the difference between the observer and random histograms, a threshold was decremented from 1 to 0, and at each threshold the percentage of eye positions and of random positions that were to a map value larger than the threshold ("hits") were computed. An ordinal dominance curve (similar to a receiver operating characteristic curve) was created with "observer-hits" versus "random-hits". The curve summarizes how well a binary decision rule based on thresholding the map values could discriminate signal (map values at observer eye positions) from noise (map values at random locations). The overall performance can be summarized by the area under the curve (AUC). An AUC area of 0.5 stands for a model which is at chance at predicting human gaze, while larger AUC values indicate better prediction performance. In our experiment, the ordinal dominance curve is plotted in Fig. 5(b), and the AUC value is $0.773 \pm 0.002$. As an upper bound, inter-observer correlations among humans yield an AUC of $0.854 \pm 0.001$.

As to the attention-based bit allocation in video compression, the latest video compression standard H.264/AVC and its reference software JM9.8 are adopted to implement the experiment. In H.264, a total of 52 different values of $Q_{step}$ are supported and they are indexed by a Quantization Parameter (QP). $Q_{step}$ increases by 12.5% for each increment of 1 in QP. In this paper, QPs are adjusted at the MB level, which means different QPs are computed for each MB. There are two reasons for this: first, in H.264 frames are encoded at the MB level, second, the generated guidance map has the same size as the frame size in MBs. QPs are computed according to Eq. (11) where $w_i$ are replaced by the corresponding GM values and $Q_{step}$ are taken from the baseline QP value. Furthermore, in order to keep the smoothness of perceptual quality, the biggest $Q_{step}$ is constrained to equal or less than 2 times of the smallest $Q_{step}$, this means the difference between QPs in one frame is constrained into 6. In the implementation, the smallest QP is set to $QP_{baseline} - 2$ while the biggest QP is set to $QP_{baseline} + 3$.

To measure the subjective quality of encoded frames, eye-tracking data are applied to compute the weighted distortion. Here we propose to use a new eye-tracking weighted mean square error (EWMSE) and eye-tracking weighted peak signal-to-noise ratio (EWPSNR) metrics

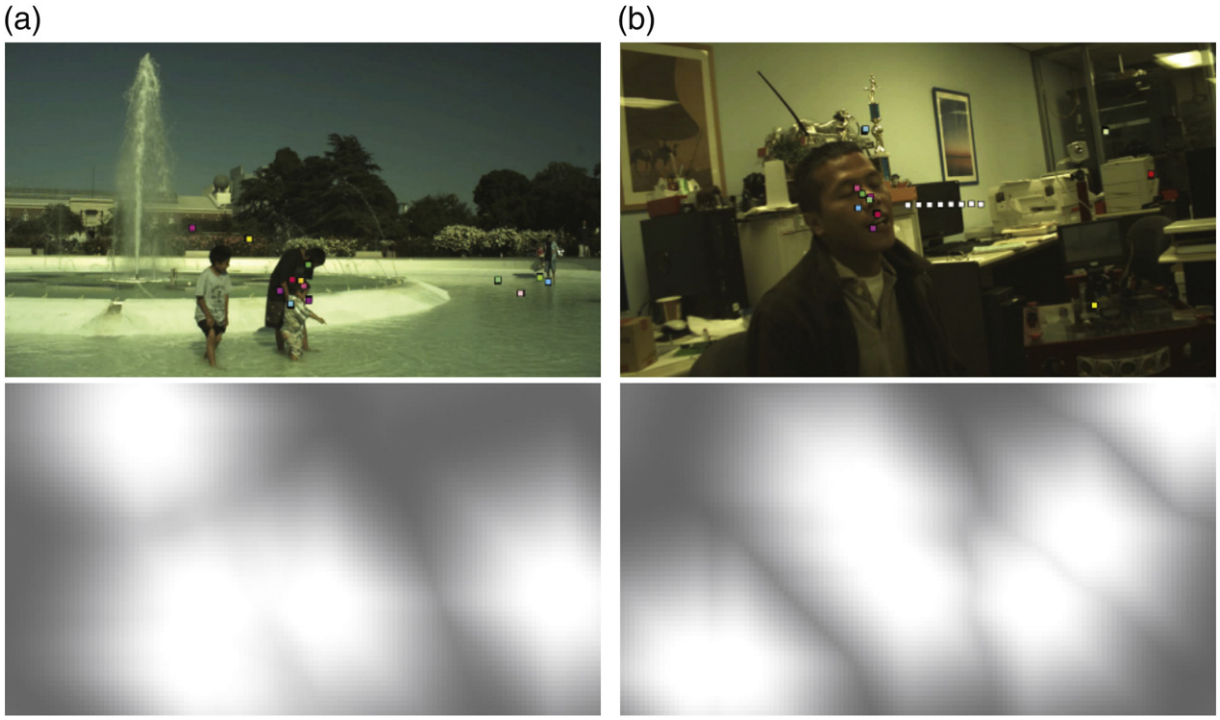



**Fig. 6.** Example of video frames (top) and the corresponding guidance maps (bottom). Subjects' eye fixations are marked as small square color patches in the frame, the white patches in (b) means a saccade. (a) Frame from *park03*, (b) frame from *room02*.

to measure subjective quality. The corresponding computation formulas are as follows:

$$EWMSE = \frac{1}{MN\sum_{x=1}^{M}\sum_{y=1}^{N} w_{x,y}} \sum_{x=1}^{M}\sum_{y=1}^{N}\left(w_{x,y}\bullet\left(I'_{x,y}-I_{x,y}\right)^2\right) \quad (12)$$

$$EWPSNR = 10^{*}\log\left(\frac{(2^n-1)^2}{EWMSE}\right) \quad (13)$$

$$w_{x,y} = \frac{1}{2\pi\sigma_x\sigma_y}e^{-\left(\frac{(x-x_e)^2}{2\sigma_x^2}+\frac{(y-y_e)^2}{2\sigma_y^2}\right)} \quad (14)$$

where $I$ and $I'$ are the original frame and the encoded frame, respectively, $M$ and $N$ are the frame's height and width in pixels, $n$ is the bit depth of the color component. $w_{x,y}$ is the weight for distortion at position $(x, y)$ and normalized to $\sum_{x,y} w_{x,y} = MN$. $w_{x,y}$ is computed based on the subjects' eye fixation position $(x_e, y_e)$ from eye-tracking experiment, $\sigma_x$and $\sigma_y$are two parameters related to the distance and view angle, usually taken from fovea size. Here we use 2° (64 pixels) of view field as $\sigma_x$ and $\sigma_y$. The rationale for this weighting formula is that the photoreceptors in the human retina are in a highly non-uniform distribution: only a small region of 2–5° of visual angle (the fovea) around the center of gaze is captured at high resolution and the resolution falls off quickly around the fovea [3]. In our experiment, the eye fixations are recorded with a 240 Hz eye-tracker, considering that the video frame rate is 30 Hz, for each subject, 8 eye

**Table 1**
Comparison of EWPSNR results between JM9.8 and the proposed model (VAGBA — visual attention guided bit allocation) for different clips. Units in the table are in dB. QP means baseline quantization parameter. Gain (in dB) is the improvement of EWPSNR compared with the standard JM9.8 method. Gain>0 means that, for the current clip, the video subjective quality encoded by the proposed method is better than the one encoded by JM9.8.

| Clip index | QP = 24 | | | QP = 28 | | | QP = 32 | | | QP = 36 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JM | VAGBA | Gain | JM | VAGBA | Gain | JM | VAGBA | Gain | JM | VAGBA | Gain |
| 1 | 41.15 | 42.89 | 1.75 | 38.66 | 40.16 | 1.50 | 35.67 | 37.49 | 1.82 | 32.77 | 34.74 | 1.97 |
| 2 | 41.36 | 42.02 | 0.66 | 38.48 | 39.27 | 0.79 | 35.41 | 36.59 | 1.18 | 32.93 | 33.93 | 1.00 |
| 3 | 43.18 | 42.99 | −0.19 | 40.56 | 40.56 | 0.00 | 38.08 | 38.17 | 0.09 | 35.77 | 35.85 | 0.08 |
| 4 | 42.67 | 42.83 | 0.16 | 40.47 | 40.52 | 0.05 | 38.03 | 38.29 | 0.27 | 34.92 | 36.08 | 1.16 |
| 5 | 41.77 | 43.08 | 1.31 | 38.93 | 40.34 | 1.41 | 36.03 | 37.58 | 1.56 | 33.21 | 34.84 | 1.64 |
| 6 | 41.95 | 42.47 | 0.52 | 39.25 | 39.74 | 0.49 | 36.54 | 37.06 | 0.51 | 33.78 | 34.46 | 0.69 |
| 7 | 42.76 | 42.69 | −0.07 | 40.06 | 40.08 | 0.02 | 37.14 | 37.49 | 0.35 | 34.32 | 34.96 | 0.65 |
| 8 | 43.35 | 43.74 | 0.39 | 40.71 | 41.20 | 0.49 | 38.35 | 38.69 | 0.34 | 35.67 | 36.18 | 0.51 |
| 9 | 42.07 | 42.58 | 0.50 | 39.29 | 39.91 | 0.61 | 36.56 | 37.30 | 0.74 | 33.97 | 34.59 | 0.62 |
| 10 | 41.51 | 42.00 | 0.49 | 38.79 | 39.24 | 0.45 | 36.06 | 36.52 | 0.46 | 33.38 | 33.84 | 0.45 |
| 11 | 41.89 | 42.50 | 0.61 | 39.34 | 39.82 | 0.48 | 36.74 | 37.23 | 0.49 | 34.04 | 34.59 | 0.56 |
| 12 | 41.30 | 41.66 | 0.36 | 38.57 | 38.88 | 0.31 | 35.73 | 36.18 | 0.45 | 33.00 | 33.51 | 0.51 |
| 13 | 42.15 | 42.76 | 0.61 | 38.98 | 40.12 | 1.14 | 36.83 | 37.50 | 0.67 | 33.68 | 34.84 | 1.16 |
| 14 | 40.22 | 41.14 | 0.92 | 37.38 | 38.23 | 0.85 | 34.57 | 35.39 | 0.82 | 31.92 | 32.66 | 0.73 |
| 15 | 41.84 | 42.57 | 0.73 | 39.14 | 40.01 | 0.87 | 36.54 | 37.42 | 0.88 | 33.92 | 34.83 | 0.91 |
| 16 | 41.22 | 42.39 | 1.17 | 38.15 | 39.56 | 1.41 | 35.48 | 36.83 | 1.36 | 32.70 | 34.11 | 1.41 |
| 17 | 43.07 | 43.45 | 0.38 | 40.44 | 40.93 | 0.49 | 37.64 | 38.45 | 0.82 | 35.24 | 35.95 | 0.71 |
| 18 | 41.29 | 42.13 | 0.84 | 38.32 | 39.48 | 1.16 | 35.64 | 36.91 | 1.27 | 33.15 | 34.40 | 1.25 |
| 19 | 41.68 | 42.67 | 0.98 | 38.72 | 39.98 | 1.26 | 36.27 | 37.32 | 1.05 | 33.49 | 34.66 | 1.17 |
| 20 | 42.73 | 42.64 | −0.09 | 40.20 | 39.99 | −0.20 | 37.55 | 37.37 | −0.18 | 34.98 | 34.75 | −0.24 |
| 21 | 42.87 | 43.39 | 0.52 | 40.25 | 40.81 | 0.56 | 37.51 | 38.23 | 0.72 | 34.91 | 35.66 | 0.75 |
| 22 | 42.99 | 42.50 | −0.48 | 40.48 | 39.92 | −0.55 | 37.90 | 37.34 | −0.56 | 35.28 | 34.81 | −0.46 |
| 23 | 44.76 | 45.70 | 0.94 | 42.57 | 43.49 | 0.92 | 40.09 | 41.09 | 1.01 | 37.65 | 38.55 | 0.90 |
| 24 | 41.69 | 42.48 | 0.79 | 39.09 | 39.83 | 0.73 | 36.44 | 37.17 | 0.72 | 33.77 | 34.51 | 0.73 |
| 25 | 43.12 | 43.55 | 0.43 | 40.61 | 41.02 | 0.41 | 38.10 | 38.47 | 0.37 | 35.49 | 35.85 | 0.36 |
| 26 | 42.68 | 42.84 | 0.16 | 39.99 | 40.12 | 0.13 | 37.18 | 37.36 | 0.18 | 34.32 | 34.62 | 0.31 |
| 27 | 44.82 | 45.50 | 0.67 | 42.75 | 43.47 | 0.72 | 40.43 | 41.20 | 0.77 | 37.94 | 38.73 | 0.79 |
| 28 | 42.45 | 43.25 | 0.81 | 39.55 | 40.76 | 1.204 | 37.17 | 38.29 | 1.12 | 35.05 | 35.79 | 0.75 |
| 29 | 42.05 | 43.10 | 1.05 | 39.05 | 40.38 | 1.33 | 36.16 | 37.68 | 1.52 | 33.41 | 34.98 | 1.57 |
| 30 | 42.18 | 42.34 | 0.16 | 39.52 | 39.58 | 0.06 | 36.93 | 36.96 | 0.03 | 34.18 | 34.38 | 0.21 |
| 31 | 43.66 | 45.95 | 2.29 | 41.72 | 43.94 | 2.22 | 38.84 | 41.61 | 2.77 | 36.34 | 39.12 | 2.77 |
| 32 | 41.15 | 42.55 | 1.40 | 38.38 | 39.77 | 1.39 | 35.28 | 37.02 | 1.74 | 32.25 | 34.28 | 2.03 |
| 33 | 42.22 | 43.04 | 0.83 | 39.48 | 40.39 | 0.91 | 36.40 | 37.79 | 1.39 | 33.86 | 35.23 | 1.37 |
| 34 | 42.69 | 43.36 | 0.67 | 40.21 | 40.79 | 0.58 | 37.72 | 38.27 | 0.55 | 34.95 | 35.84 | 0.88 |
| 35 | 44.65 | 45.38 | 0.74 | 43.08 | 43.61 | 0.54 | 41.01 | 41.48 | 0.47 | 38.75 | 39.21 | 0.46 |
| 36 | 42.15 | 42.68 | 0.53 | 39.17 | 40.01 | 0.84 | 36.35 | 37.38 | 1.03 | 33.33 | 34.79 | 1.46 |
| 37 | 42.49 | 42.84 | 0.35 | 39.73 | 40.13 | 0.40 | 36.97 | 37.50 | 0.53 | 34.26 | 34.90 | 0.64 |
| 38 | 42.56 | 43.37 | 0.81 | 40.08 | 40.82 | 0.74 | 37.54 | 38.27 | 0.73 | 34.85 | 35.71 | 0.87 |
| 39 | 42.00 | 42.04 | 0.03 | 39.49 | 39.36 | −0.13 | 36.92 | 36.77 | −0.15 | 34.42 | 34.22 | −0.20 |
| 40 | 41.77 | 42.79 | 1.02 | 38.74 | 40.08 | 1.35 | 35.74 | 37.46 | 1.72 | 32.95 | 34.84 | 1.89 |
| 41 | 42.73 | 43.17 | 0.44 | 39.79 | 40.48 | 0.69 | 37.00 | 37.88 | 0.88 | 34.43 | 35.30 | 0.87 |
| 42 | 41.86 | 42.62 | 0.76 | 39.29 | 39.98 | 0.69 | 36.74 | 37.40 | 0.66 | 34.23 | 34.91 | 0.68 |
| 43 | 43.61 | 44.05 | 0.44 | 40.82 | 41.54 | 0.72 | 38.22 | 39.02 | 0.79 | 35.64 | 36.56 | 0.93 |
| 44 | 42.79 | 43.67 | 0.88 | 40.46 | 41.12 | 0.65 | 38.59 | 38.53 | −0.06 | 35.80 | 36.03 | 0.23 |
| 45 | 42.79 | 42.84 | 0.05 | 40.22 | 40.24 | 0.02 | 37.41 | 37.63 | 0.223 | 34.51 | 35.08 | 0.56 |
| 46 | 41.63 | 42.74 | 1.10 | 38.97 | 39.97 | 1.00 | 36.26 | 37.26 | 1.01 | 33.37 | 34.59 | 1.22 |
| 47 | 42.14 | 43.64 | 1.50 | 39.23 | 41.05 | 1.81 | 36.74 | 38.47 | 1.73 | 34.17 | 35.95 | 1.79 |
| 48 | 42.51 | 42.74 | 0.23 | 39.89 | 40.14 | 0.25 | 37.15 | 37.50 | 0.36 | 34.27 | 34.90 | 0.62 |
| 49 | 41.54 | 43.07 | 1.53 | 38.65 | 40.31 | 1.66 | 35.58 | 37.58 | 2.01 | 32.57 | 34.83 | 2.26 |
| 50 | 42.48 | 43.81 | 1.33 | 40.16 | 41.31 | 1.15 | 37.02 | 38.77 | 1.75 | 34.42 | 36.22 | 1.79 |
| Average | 42.36 | 43.04 | 0.68 | 39.71 | 40.44 | 0.73 | 37.03 | 37.85 | 0.82 | 34.35 | 35.27 | 0.92 |

fixation points need to be taken into account for each frame. Therefore, the weight $w_{x,y}$ in reality is a combination of all 8 different eye fixation points. Furthermore, the saccade data are not considered in computing the EWPNSR and only the fixation points are taken into account. We did this because human take saccade very quickly and do not pay much attention to the saccade regions. The mean EWPSNR from all the subjects is adopted as the measurement to evaluate the video subjective quality: the higher EWPNSR value, the better subjective quality.

To show the effectiveness of the proposed visual attention guided bit allocation method in improving the video subjective quality, we compare the encoded video EWPSNR from the proposed method and the standard method in JM9.8 with matched bit rate through the frame-level rate control algorithm. The configuration of the encoder is as follows: intra period = 30, Hadamard transform, UVLC, no fast motion estimation, no B frame, high complexity RDO mode, no restriction in search range. We test 4 baseline QPs (initial QP): 24, 28, 32 and 36, and the bit rates range from 260 Kbps to 10 Mbps. The rate-controlled bit rates with the standard encoder precisely match the bit rate with the proposed new encoder (within 1% difference). Table 1 lists all the EWPSNR results from proposed method, the results from the JM9.8 standard method, and the subjective quality improvement (gain). Better EWMSE and EWPSNR is expected to be obtained for our method vs. JM only if, on average, the predictions of the saliency model agree with where humans look, and, when they disagree, the higher distortions that our system will introduce at the locations looked at by humans do not outweigh the lower distortions obtained when the model is correct. In addition, Fig. 7 plots the results at different baseline QP, and sorts the results according to improvement. It is easy to see that only for a few (3–4) clips the subjective quality is worse with our proposed method than with the standard method, while most of the clips achieve a better subjective quality, with improvement for some of them up to about 2 dB EWPSNR. Thus, the proposed scheme can significantly ($p<0.002$, $t$-test) achieve better subjective quality (as defined by our EWPSNR measure) while keeping the same bit rate. Furthermore, the comparison of histograms of PSNR results at eye fixation regions (2° of visual field) with different methods is plotted in Fig. 8, from the figure it is easy to see that more encoded frames have higher PSNR in the fixation region with the proposed method compared with the standard JM9.8 method. Fig. 9 shows two examples of EWPSNR over the clip frames. We can see

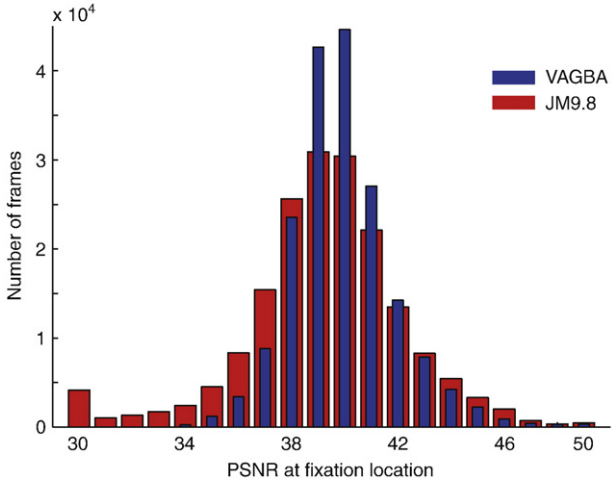



**Fig. 8.** Comparison of histograms of PSNR results at eye fixation regions with standard JM9.8 method and the proposed VAGBA method (initial QP = 28). The fixation regions are used 2° of view.

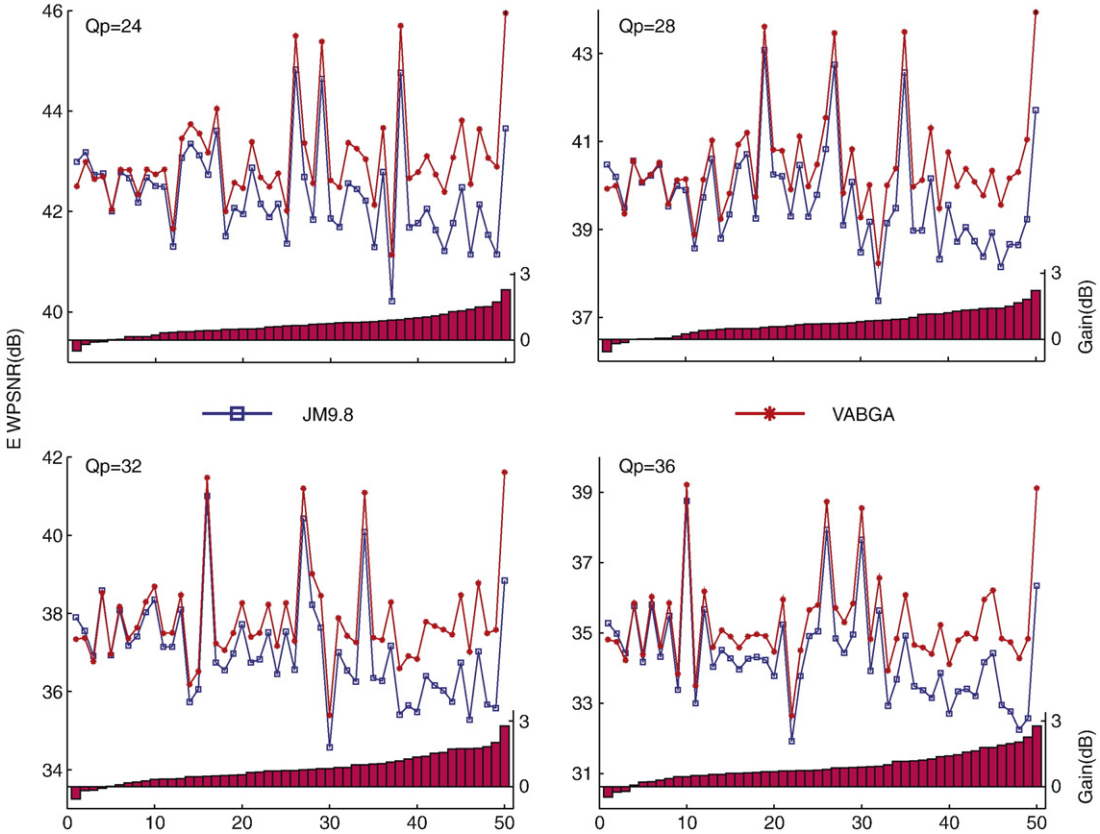



**Fig. 7.** EWPSNR results comparison between our proposed VAGBA model and JM9.8 at the same bit rate, the horizontal axis represent the clip index, after sorting by the subjective quality improvement (the red bars shown in each plot).

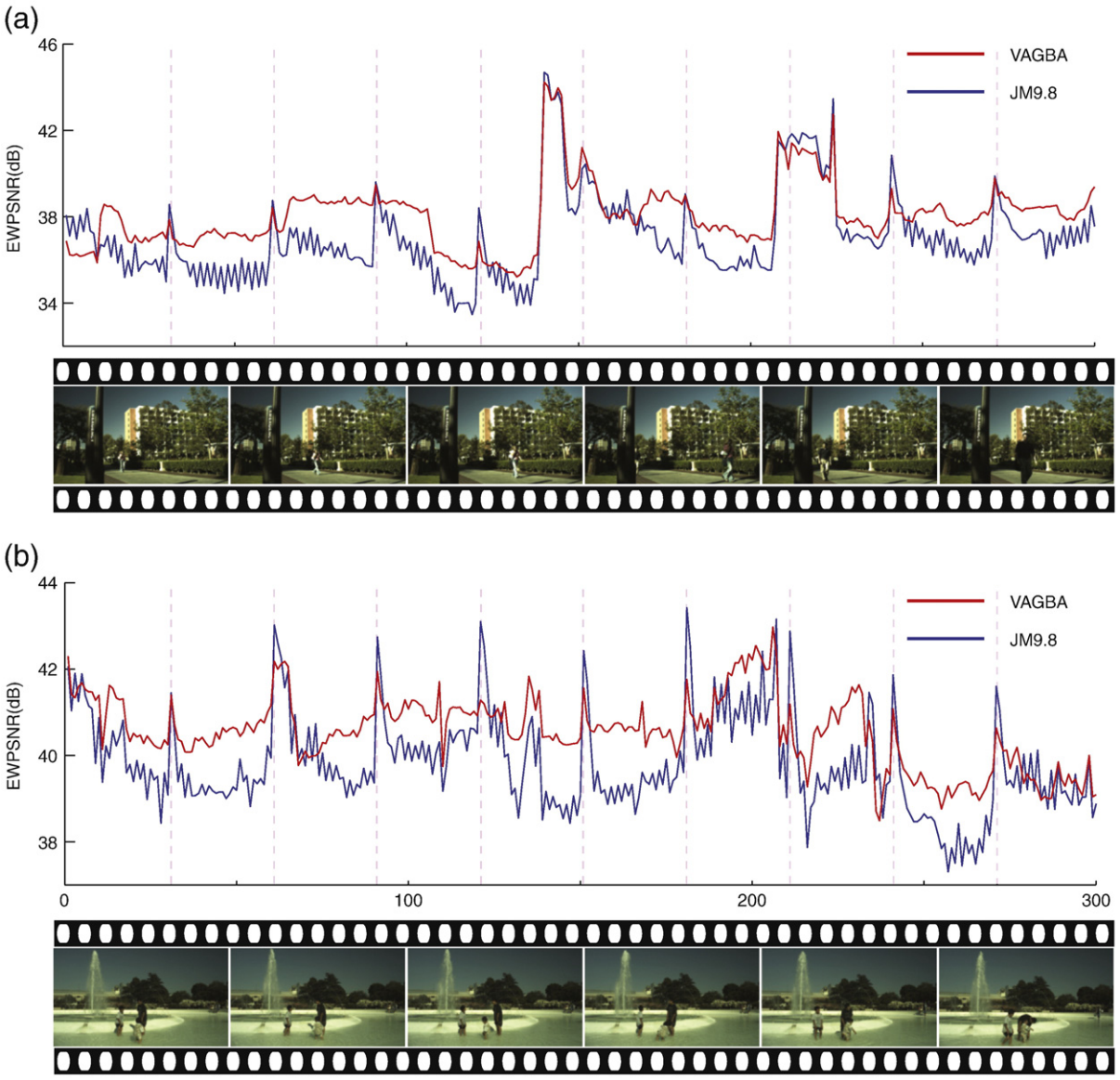



**Fig. 9.** Two examples of EWPSNR over the clip frames. (a) *garden09* from subject KC, initial QP = 28; (b) *park03* from subject SH, initial QP = 28. Dashed grey lines indicate intra-coded frames (every 30 frames with our codec settings).

from the figure that for most frames, the subjective quality of the proposed method is better than the standard JM9.8 encoded result. However, for some frames, the standard method achieves better subjective quality. There are mainly two reasons for this: first, if the current frame is an intra refresh frame, the rate control algorithm in H.264 usually assigns a relatively smaller QP to such intra frame to achieve better prediction results for later P frames. In these cases, the current frame's quality could be better than with the proposed method. Second, the attention prediction model is not guaranteed to always accurately predict human's attention regions for all the frames, such that if the prediction failed for the current frame, then due to the bit allocation strategy, the true attention region will receive fewer bits to encode thus will make the subjective quality worse.

Furthermore, the comparison between the proposed method and our previous method proposed in [21] which guide video compression through selective blurring (foveation) of low-salience image regions is conducted. The foveated clips are encoded by JM9.8 with matched bit rate through the frame-level rate control algorithm (within 1% difference). Fig. 10 shows the comparison results at different baseline QP, and sorts the results according to improvement of EWPSNR. From the figure we can see that for all clips the subjective quality is significantly better with the proposed method than with the foveation method ($p<10^{-10}$, $t$-test). The average improvement in EWPNSR is 2.533 dB. Also we can see from the figure that the improvement is higher when the baseline QP is lower, this is because the foveation degrade the video quality more than the encode error when the quantization step is small. As another example, Fig. 11 compares the visual qualities of three partially reconstructed frames among the standard rate control method, the foveation method and the proposed bit allocation method. The difference frames (encoding error) are also plotted to make the comparison clearer. As shown in the figure, the encoded frame by the proposed method has better visual quality than the frame encoded by standard method in the interesting region.

## 6. Discussion

The proposed attention model predicts human attention accurately in most cases, and based on this, the bit allocation algorithm can improve the subjective visual quality significantly while keeping the same bit rate. The contributions in this paper include several aspects:

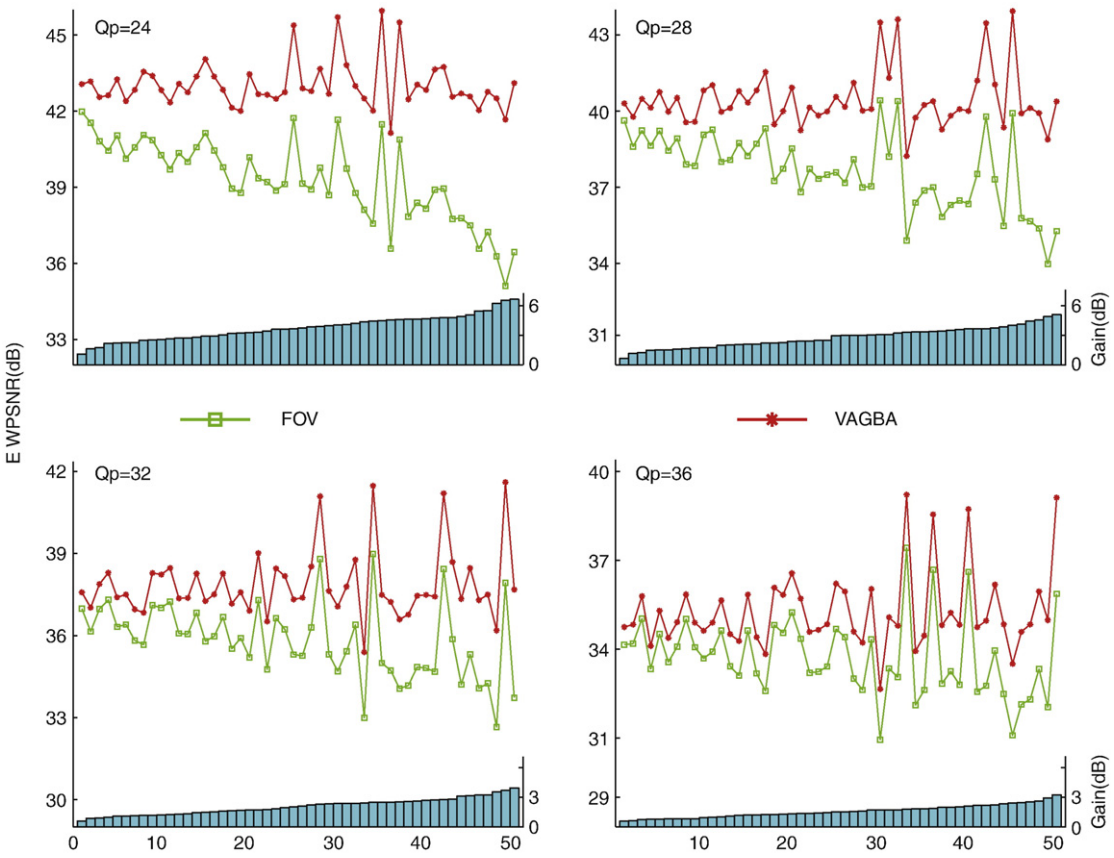
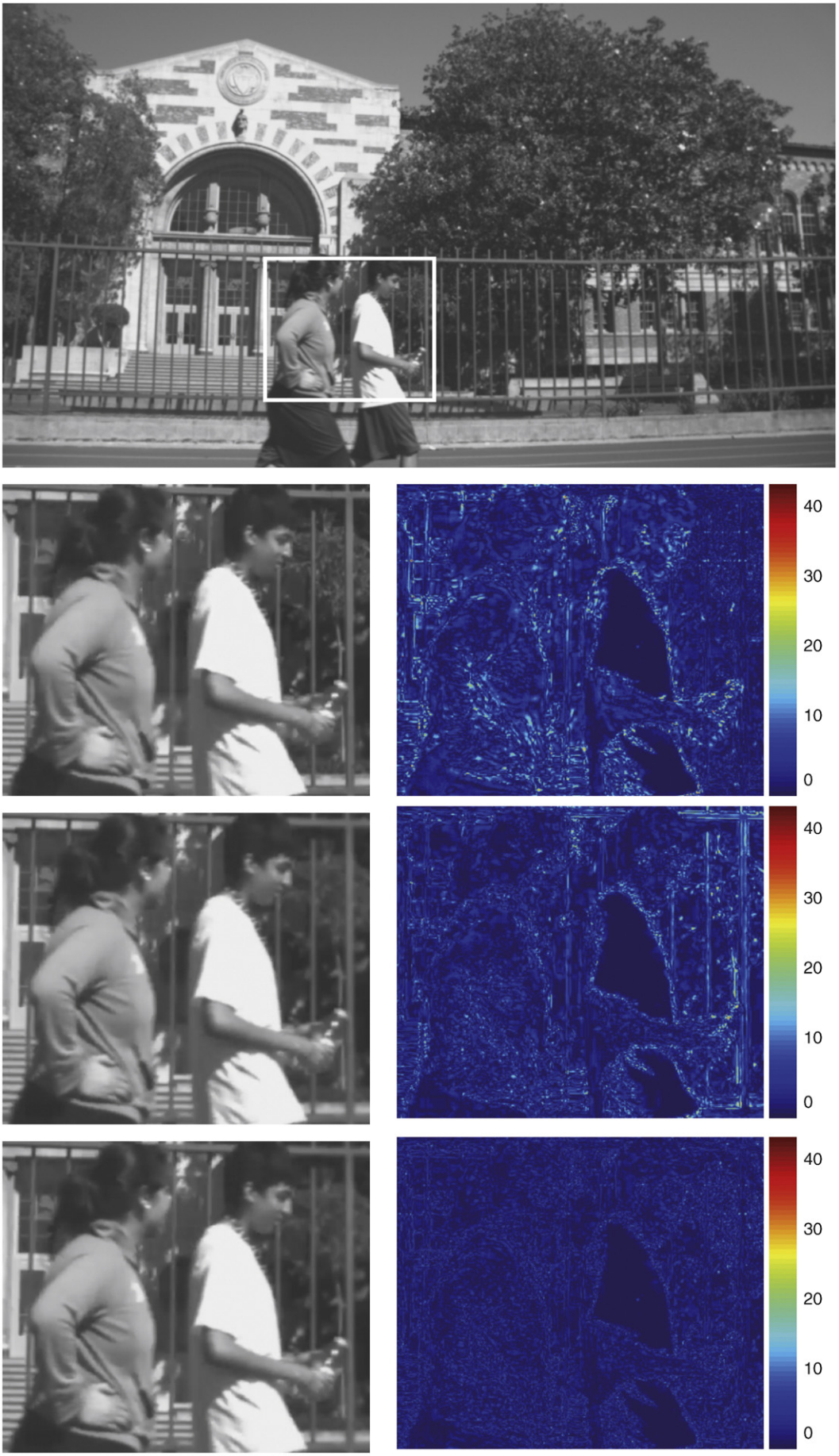



**Fig. 10.** EWPSNR results comparison between our proposed VAGBA model and the foveation method (mark as FOV in the figure) at the same bit rate, the horizontal axis represent the clip index, after sorting by the subjective quality improvement (the green bars shown in each plot).

(1) Combining the attention prediction model with state-of-the-art video compression method, the proposed method is fully automatic and can be applied to any kind of video categories without any restriction. In addition to combining with the H.264 codec, the proposed method can be applied to any kind of codec proposed so far. (2) A new bit allocation strategy was proposed through solving in closed form the constrained global optimization problem. (3) Using eye-tracking data to evaluate compressed video frame quality in a quantitative way (instead of subjective rating or binary selection). The new proposed EWPSNR subjective quality measurement is based on the human vision's characteristic and can represent the non-uniform subjective distortion in a reasonable way. (4) Collecting a high-definition video sequence database with eye-tracking data and distributing them on the Internet. This dataset can be used for video compression purposes as well as attention prediction purposes. Also, the raw captured frames are in Bayer format, which means this dataset can be used for Bayer format related image processing research.

The target in this paper is to validate the effectiveness in improving the subjective quality while keeping the same bit rate and employing a purely algorithmic method which does not require manual parameter tuning. Thus far, the bit allocation strategy is an open-loop algorithm, which means that it only adjusts the bit allocation according to the guidance map and takes no constrain to keep any presumed bit rate. In our implementation, we first compress the video sequences with the proposed method, after that, we use the available rate control algorithm in JM9.8 to match the bit rate and use this result to compare with our proposed method. The comparison is reasonable because there is no scene change in the test sequences and thus both bit rate and visual quality should not fluctuate too much over video frames. Although the proposed method is not bit rate constrained, it can be used in many applications such as video storage, band-free video stream transmission, etc. However, it remains a great task to develop a bit rate constrained bit allocation strategy based on the visual attention model. Another point which is not addressed by an open-loop algorithm but could be studied more closely in the future is how the algorithm itself may introduce artifacts in the low-bit rate regions of the compressed video, which may themselves be salient and attract human attention (see [21] for more detailed discussion). This could be addressed in future versions of our algorithm where the saliency map may be computed on the compressed video clips as well, to check for the introduction of possibly salient artifacts during compression.

Eye-tracking data recorded from subjects viewing the uncompressed video clips are applied in evaluating the subjective quality. The rationale for viewing uncompressed video lies in two aspects: first, the eye-tracking traces from the uncompressed videos show the real attention regions of the original clips. Second, the ideal subjective quality measurement should use eye-tracking data from both the proposed method encoded video and standard rate-controlled video. However, it is impossible to obtain these two kinds of eye-tracking data from the same subject without affecting a priori knowledge: no matter which kind of encoded video is presented to subjects first, the eye-tracking data from the second presentation would likely be affected by the fact that subjects have already seen essentially the same clips before. Considering that artifacts might attract attention, two steps are adopted to reduce the quality fluctuation: first, both temporal and spatial smooth operation are conducted in computing the guidance map; Second, the biggest $Q_{step}$ is constrained to equal or less than 2 times of the smallest $Q_{step}$ in one frame to keep the perceptual quality, in the implementation, the smallest QP is set to $QP_{baseline}-2$ while the biggest QP is set to


40
30
20
10
0
40
30
20
10
0
40
30
20
10
0

$QP_{baseline} + 3$. In our experiment, we check the compressed videos and found there is no big quality fluctuation in both spatial and temporal.

EWPSNR is proposed in computing the subjective quality. Here, we wanted to investigate whether we could test our algorithm in a more objective and more informative way than using subjective quality ratings, where observers may not always be able to rationalize or explain their ratings. Many existing computational subjective quality measurements [26,41,42] often tend to rely on some knowledge about the human visual system to decide what may be more visible or important to a human observer. For example, see the measures of FPNSR (Foveated PSNR, [26]), SSIM (structural similarity, [41]), and DVQ (Digital Video Quality, [42]). Thus we have been concerned that using these measures may be somewhat circular: (1) process video images with some saliency algorithm and allocate more bits (lower distortion) to more salient regions; (2) measure subjective quality with an algorithm that is very similar to our saliency computation and hence may be quite strongly correlated with it. We would quite naturally expect good subjective quality. This is what prompted us to develop the EWPSNR metric. We believe that it is an objective way to measure subjective quality, and it has the advantage of not relying on any algorithmic assumptions regarding how subjective quality may be defined. Here we just assume that the locations which people look at are the ones which will matter in terms of subjective quality. Note that this assumption is itself an imperfect one (subjective quality seems to be influenced not only by foveal vision but also by peripheral vision). But we believe that it at least avoids circularity in our testing, and it also provides a very informative assessment of where the algorithm is working (good agreement between the algorithm and human gaze) or failing.

Over the 50 tested video clips, there are 3–4 cases in which the subjective quality of clips encoded by our proposed method is worse than the clips encoded by the standard H.264 method. The worst two clips are *gate03* and *seagull01*, example frames from these two clips can be seen in Figs. 2 and 4. The reason of the failure mainly is that, for these clips, the attention prediction model results do not match well the subjects' attention. In the proposed attention prediction model, high motion regions take higher saliency value, however, in the *gate03* clip, the high speed cars were less interesting to our human observers than the jogging girl and the flags. In *seagull01* clip, the seagulls fly everywhere and the video is less meaningful in content, the subjects' eye-tracking traces are highly divergent, thus the proposed attention prediction model cannot predict the attention for all the subjects accurately.

The proposed attention prediction model in this paper purely depends on the bottom-up low-level features. These features are independent of the video contents and can be applied to any kind of conditions. However, in many specific cases, top-down influences can be taken into consideration to improve the attention prediction performance. For example, in teleconference videos or face-oriented conditions, face information (using, e.g., face detection algorithms) can be added as an important factor in predicting the attention [11]. In a specific search task (person detection), the saliency, target features, and scene context combined models can predict 94% of human agreement [43]. Also, considering the layout information, the "gist" of scenes can be applied to improve the prediction by learning broad scene categories [44,45]. All these top-down factors may combine with the bottom-up model to improve the attention prediction performance and thus improve the subjective quality of compression in specific corresponding conditions.

## Acknowledgments

This study was supported by NSF, ARO, DARPA, Natural Science Foundation of China (NSFC) and China Scholarship Council (CSC). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

## References

[1] T. Weigand, G.J. Sullivan, A. Luthra, Draft ITU-T recommendation H.264 and final draft international standard 14496-10 advanced video coding, Joint Video Teams of ISO/IECJTCI/SC29/WG11 and ITU-T SG/16/Q.6Doc.JVT-G050r, Geneva, Switzerland, 2003, May.
[2] T. Wiegand, G.J. Sullivan, G. Bjntegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits and Syst. Video Technol, vol. 13, July 2003, pp. 560–576.
[3] B. Wandell, Foundations of Vision, Sinauer, Sunderland, MA, 1995.
[4] P.T. Kortum, W.S. Geisler, Implementation of a foveated image coding system for bandwidth reduction of video images, Proc. SPIE, vol. 2657, 1996, pp. 350–360.
[5] N. Doulamis, A. Doulamis, D. Kalogera, S. Kollias, Improving the performance of MPEG coders using adaptive regions of interest, IEEE Trans. On Circuits syst. Video Technol, vol. 8, 1998, pp. 928–934.
[6] S. Lee, A.C. Bovik, Y.Y. Kim, Low delay foveated visual communications over wireless channels, Proc. IEEE Int. Conf. Image Processing, 1999, pp. 90–94.
[7] U. Rauschenbach, H. Schumann, Demand-driven image transmission with levels of detail and region s of interest, Comput. Graph, vol. 23, no. 6, 1999, pp. 857–866.
[8] D.J. Parkhurst, E. Niebur, Variable-resolution displays: a theoretical, practical, and behavioral evalutation, Human Factors, vol. 44, no. 4, 2002, pp. 611–629.
[9] M. Chi, M. Chen, C. Yeh, J. Jhu, Region-of-interest video coding based on rate and distortion variations for H.263+, Image Communication, vol. 23, no. 2, 2008, pp. 127–142.
[10] O. Hershler, S. Hochstein, At first sight: a high-level pop out effects for faces, Vision Research, vol. 45, no. 13, 2005, pp. 1707–1724.
[11] M. Cerf, J. Harel, W. Einhauser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, Advances in neural information processing systems, vol. 20, 2008, pp. 241–248.
[12] K.C. Lai, S.C. Wong, K. Lun, A rate control algorithm using human visual system for video conferencing systems, Proc. Int. Conf. Signal Processing, vol. 1, Aug. 2002, pp. 656–659.
[13] L. Tong, K.R. Rao, Region-of-interest based rate control for low-bit-rate video conferencing, Journal of Electronic Imaging, vol. 15, no. 3, July, 2006.
[14] L. S. Karlsson, "Spatio-temporal pre-processing methods for region-of-interest video coding," PhD dissertation, Sundsvall, Sweden, 2007.
[15] C.-W. Tang, C.-H. Chen, Y.-H. Yu, C.-J. Tsai, Visual sensitivity guided bit allocation for video coding, IEEE Trans. Multimedia, vol. 8, Feb. 2006, pp. 11–18.
[16] K. Minoo, T.Q. Nguyen, Perceptual video coding with H.264, Proc. 39th Asilomar Conf. Signals, Systems, and Computers, Nov. 2005.
[17] C. Huang, C. Lin, A novel 4-D perceptual quantization modeling for H.264 bit-rate control, IEEE Trans. On Multimedia, vol. 9, no. 6, 2007, pp. 1113–1124.
[18] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, Nov. 1998, pp. 1254–1259.
[19] L. Itti, C. Koch, Computational modeling of visual attention, Nature Reviews, Neuroscience, vol. 2, Mar, 2001, pp. 194–203.
[20] T. Liu, N. Zheng, W. Ding, Z. Yuan, Video attention: learning to detect a salient object sequence, Proc. ICPR, 2008, pp. 1–4.
[21] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. on Image Processing, vol. 13, no. 10, 2004, pp. 1304–1318.
[22] L. Itti, Automatic attention-based prioritization of unconstrained video for compression, Proc. SPIE Human Vision and Electronic Imaging, vol. 5292, 2004, pp. 272–283.
[23] W. Lai, X. Gu, R. Wang, W. Ma, H. Zhang, A content-based bit allocation model for video streaming, Proc. IEEE international Conference on Multimedia and Expo, 2004, ICME).
[24] Z. Chen, G. Qiu, Y. Lu, L. Zhu, Q. Chen, X. Gu, W. Charles, Improving video coding at scene cuts using attention based adaptive bit allocation, Proc. ISCAS, 2007, pp. 3634–3638.
[25] M. Jiang, N. Ling, On Lagrange multiplier and quantizer adjustment for H.264 frame layer video rate control, IEEE Trans. Circuits and Syst. Video Technol, vol. 16, no. 5, 2006, pp. 663–669.
[26] S. Lee, M.S. Pattechis, A.C. Bovik, Foveated video compression with optimal rate control, IEEE Trans. on Image Processing, vol. 10, no. 7, 2001, pp. 977–992.
[27] Y. Sun, I. Ahmad, D. Li, Y. Zhang, Region-based rate control and bit allocation for wireless video transmission, IEEE Trans. on Multimedia, vol. 8, no. 1, 2006, pp. 1–10.

**Fig. 11.** Example of visual quality comparison between encoded frame (partially) by proposed method, foveation method and standard method. Top part is the Y component of the original 230th frame from *field03* clip. From row 2 to row 4 are the results from JM9.8 rate control method (PSNR = 38.37 dB), foveation method (PSNR = 30.76 dB), and proposed method (PSNR = 39.37 dB), respectively. For each row from row 2 to row 4, left side shows the encoded frame while right side shows the encode error. All the encoded frames are shown only partially to the interesting region.

[28] Y. Liu, Z. Li, Y.C. Soh, Region-of-interest based resource allocation for conversational video communications of H.264/AVC, IEEE Trans. Circuits and Syst. Video Technol, vol. 18, no. 1, 2008, pp. 134–139.
[29] Z. Li, L. Itti, Visual attention guided video compression, Proc. Vision Science Society Annual Meeting (VSS08), May, 2008.
[30] N. Wang, Q. Zhang, G. Li, Objective quality evaluation of digital video, Proc. PCCAS, 2000, pp. 791–794.
[31] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, Objective video quality assessment system based on human perception, Proc. SPIE, vol. 1913, 1993, pp. 15–26.
[32] A.B. Watson, J. Hu, J.F. McGowan, Digital video quality metric based on human vision, Journal of Electronic Imaging, vol. 10, no. 1, 2001, pp. 20–29.
[33] "Tutorial: objective perceptual assessment of video quality: full reference television,", ITU-T Technical tutorials, 2005.
[34] http://ilab.usc.edu/toolkit.
[35] J.M. Wolfe, Visual memory: what do you know about what you saw? Current Biology, vol. 8, 1998, pp. 303–304.
[36] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognition Psychology, vol. 12, 1980, pp. 97–136.
[37] L. Itti, C. Koch, Feature combination strategies for saliency-based visual attention systems, Journal of Electronic Imaging, vol. 10, Jan. 2001, pp. 161–169.
[38] Y. Wang, J. Ostermann, Y. Zhang, Video Processing and Communication, Pearson Education, 2002.
[39] H.S. Malvar, L.W. He, R. Cutler, High-quality linear interporlation for demosaicing of bayer-patterned color images, Proc. ICASSP, 2004, pp. 485–488.
[40] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, Journal of Mathematical Psychology, vol. 12, 1975, pp. 375–387.
[41] M. Vranjes, S. Rimac-Drlje, D. Zagar, Subjective and objective quality evaluation of the H.264/AVC coded video, Proc. Systems, Signals and Image Processing, 2008.
[42] A.B. Watson, Toward a perceptual video quality metric, Human Vision, Visual Processing, and Digital Display, vol. 3299, 1998, pp. 139–147.
[43] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, A. Oliva, Modeling search for people in 900 scenes: a combined source model of eye guidance, Visual Cognition, vol. 17, no. 6&7, 2009, pp. 945–978.
[44] Z. Li, L. Itti, Gist based top-down templates for gaze prediction, Proc. Vision Science Society Annual Meeting, 2009, VSS09.
[45] J. Peters, L. Itti, Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention, Proc. CVPR, 2007.

# Saliency and Gist Features for Target Detection in Satellite Images

Zhicheng Li and Laurent Itti

***Abstract*—Reliably detecting objects in broad-area overhead or satellite images has become an increasingly pressing need, as the capabilities for image acquisition are growing rapidly. The problem is particularly difficult in the presence of large intraclass variability, e.g., finding "boats" or "buildings," where model-based approaches tend to fail because no good model or template can be defined for the highly variable targets. This paper explores an automatic approach to detect and classify targets in high-resolution broad-area satellite images, which relies on detecting statistical signatures of targets, in terms of a set of biologically-inspired low-level visual features. Broad-area images are cut into small image chips, analyzed in two complementary ways: "attention/saliency" analysis exploits local features and their interactions across space, while "gist" analysis focuses on global nonspatial features and their statistics. Both feature sets are used to classify each chip as containing target(s) or not, using a support vector machine. Four experiments were performed to find "boats" (Experiments 1 and 2), "buildings" (Experiment 3) and "airplanes" (Experiment 4). In experiment 1, 14 416 image chips were randomly divided into training (300 boat, 300 non-boat) and test sets (13 816), and classification was performed on the test set (ROC area: $0.977 \pm 0.003$). In experiment 2, classification was performed on another test set of 11 385 chips from another broad-area image, keeping the same training set as in experiment 1 (ROC area: $0.952 \pm 0.006$). In experiment 3, 600 training chips (300 for each type) were randomly selected from 108 885 chips, and classification was conducted (ROC area: $0.922 \pm 0.005$). In experiment 4, 20 training chips (10 for each type) were randomly selected to classify the remaining 2581 chips (ROC area: $0.976 \pm 0.003$). The proposed algorithm outperformed the state-of-the-art SIFT, HMAX, and hidden-scale salient structure methods, and previous gist-only features in all four experiments. This study shows that the proposed target search method can reliably and effectively detect highly variable target objects in large image datasets.**

***Index Terms*—Gist features, saliency features, satellite images, target detection.**

Manuscript received August 05, 2010; revised November 19, 2010; accepted November 24, 2010. Date of publication December 13, 2010; date of current version June 17, 2011. This work was supported by Defense Advanced Research Projects Agency under Government Contract HR0011-10-C-0034, the National Geospatial Intelligence Agency under Grant 19-1082141, the National Science Foundation under CRCNS Grant BCS-0827764, the Army Research Office under Grant W911NF-08-1-0360, and the China Scholarship Council under Grant 2007103281. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kenneth K. M. Lam.

Z. Li was with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191 China. He is now with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: lzcbuaa@gmail.com).

L. Itti is with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: itti@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2099128

## I. Introduction

OVERHEAD and satellite imagery have become ubiquitous, with applications ranging from intelligence gathering to consumer mapping and navigation assistance. With the overwhelming amount of satellite imagery available today, it has become impossible for human image analysts to examine all of the imagery, in search of interesting intelligence information. Thus, there is a pressing need for automatic algorithms to preprocess the data and to extract actionable intelligence from raw imagery, thereby facilitating and supporting human interpretation. This paper focuses on automatically detecting diverse types of targets with large intraclass variability in satellite images. This analysis is one of the currently highly time-consuming tasks that image analysts routinely perform manually. Providing new means to automate this task is expected to facilitate and render more efficient the interpretation of satellite image by human analysts.

The problem of target detection is a difficult challenge in computer vision [1]–[3]. For a given scene (image), the target detection task can be simply described as "where is the target?" Considering the feature types used for detection in static images, algorithms for target detection can be briefly summarized as belonging to three broad categories: A first, relatively straightforward approach is to use a provided (or trained) target template or model (hence, the feature is the image itself), to match against targets in the image of interest, at different locations, orientation and scales [4]–[6]. This type of method works well when the variability of targets is small (for example, detecting human faces [5], [6]). A second method for target detection is to use a model to extract a spatially sparse collection of invariant structural features (e.g., keypoint descriptors, bags of features) of the target even when viewpoint, pose, and lighting conditions vary [7]–[10]. In a third approach, using knowledge of target shape and characteristic geometry, several studies have proposed methods which learn and apply target geometric constraints on the keypoint feature locations [11], [12]. In practice, the detection algorithms usually overlap these categories, and some approaches are intermediate between the geometry-based and "bag of features" approaches retaining only some coarsely-coded location information or recording the locations of features relative to the target's center [3], [13]. In addition to these machine vision approaches, several biologically-inspired computational models have also started exploring target detection tasks in imagery, usually based on our knowledge of visual cortex, showing some promising experimental results [14]–[19]. Our approach extends these biologically-inspired frameworks.

Based on the special properties of satellite image, several algorithms have been proposed to detect the targets in such kind

1057-7149/$26.00 © 2010 IEEE

of images. For example, for hyper-spectral satellite images, the features applied usually take advantage of the reflection characteristics of different materials [20]–[22] while for multispectral images, the features are usually extracted from fused spectra [23], [24]. However, the images discussed in this paper focus on the visible spectrum and, thus, the detection methods discussed in the previous paragraph are usually adopted. Despite all the recent advances in computer vision technologies, humans still perform orders of magnitude better than the best available vision systems in object and target detection, and for many target search applications humans remain the gold standard. As such, it is reasonable to examine the low-level mechanisms as well as the system-level computational architecture of human vision for inspiration. Early on, the human visual processing system already makes decisions to focus attention and processing resources onto those small regions within the field of view which look more interesting or visually "salient" [25]–[27]. When no specific search target, no search task, and no particular time or other constraint are specified to an observer, bottom-up (image-derived) information may play a predominant role in guiding attention toward potential generically interesting targets [28]. The mechanism of selecting a small set of candidate salient locations in a scene has recently been the subject of comprehensive research efforts and several computational models have been proposed [29]–[34]. One can make use of these models to predict possible target locations and target distributions. In this paper, saliency maps from several feature channels (intensity contrast, local edge orientation, etc.) are computed from a modified Itti-Koch saliency model [25], [31], [35]. Given a static or dynamic visual scene, this model creates a number of multiscale topographic feature maps which analyze the visual inputs along visual feature channels known to be represented in the primate brain [31] and thought to guide visual attention and search [36] (luminance contrast, color-opponent contrast, oriented edges, etc.). Center-surround mechanisms and long-range competition for salience operate separately within each feature channel, coarsely reproducing neuronal interactions within and beyond the classical receptive field of early sensory neurons [37], [38]. These interactions are critical in transforming raw feature responses (e.g., an edge map computed over the input scene) into salient feature responses, as they emphasize locations which are locally outliers to the global statistics of the scene. As a result, local feature responses (e.g., a color contrast response to a small red object in an image) are modulated globally depending on the entire scene's content (e.g., the response to the small red object might be inhibited if many other red objects are present in the scene, or might be amplified if all other objects in the scene are blue). After these interactions, the feature maps from all feature channels are combined into a single scalar topographic saliency map. Locations of high activity in the saliency map are more likely to attract attention and gaze [28], [29].

Thus far, saliency-based analysis of scenes has been predominantly applied to relatively small images, typically on the order of 1 megapixel (MP), with at least one study pushing to 24 MP [40]. Such smaller images are coarsely matching the amount of information which might arise from a primate retina (about 1 million distinct nerve fibers in each of the human optic nerves). With larger broad-area-search images, for example 400 MP–1000 MP satellite images, it becomes an interesting research question whether the mechanisms developed by the primate brain might scale up. Here, we address this question by developing a new algorithm, which analyzes large images in small chips, thus, mimicking the processing which human image analysts might operate when they deploy multiple eye fixations on an image, analyzing each fixated location in turn. A second important research question is whether saliency maps might be useful at all for object classification, as opposed to being limited to just attention guidance as described previously. Here we hypothesize that, within each chip, the chip's saliency map may provide a coarse indication of the structure of the visual contents of the chip. Hence, rather than attempting to shift an attention spotlight to different salient locations within the chip, the hypothesis underlying the proposed algorithm is that a coarse analysis of the statistics of a chip's saliency map may provide sufficient clues for classifying the chip as containing or not a target. For example, target chips might have more numerous and sharper saliency peaks than nontarget chips. Our experiments and results test whether this approach is viable for complex target classification tasks where the intraclass heterogeneity is significant (e.g., find "boats," ranging from small pleasure craft to larger commercial or military ships). For each saliency map, mean, variance, number of local maxima, and average distance between the locations of local maxima are adopted to summarize saliency maps. These values to some extent represent the saliency intensity and the salient objects' spatial distribution. In the full algorithm described in the following, all of these values from different feature channels' saliency maps are combined together to form the "saliency features" part of the proposed algorithm.

Parallel with attention guidance and mechanisms for saliency computation, studies of scene perception have shown that observers can recognize the "gist" of a real-world scene from a single, possibly very brief glance. For example, following presentation of a photograph for just a fraction of a second, a human observer may report that it is an indoor meeting room or an outdoor scene of a beach [41]–[45]. Such a report from the first glance onto an image is remarkable considering that it summarizes the quintessential characteristics of an image, a process previously thought to require deep visual and cognitive analysis. With very brief exposures (100 ms or below), reports are typically limited to a few general semantic attributes (e.g., indoors, outdoors, playground, mountain) and a coarse evaluation of the distributions of visual features (e.g., grayscale, colorful, large masses, many small objects) [46]–[48]. Gist may be computed in brain areas which have been shown to preferentially respond to "places," that is, visual scene types with a restricted spatial layout [49]. Like Siagian-Itti's gist formulation in computer vision [50], here we use the term "gist" to represent a low-dimensional (compared with the raw image pixel array) scene representation feature vector which is acquired over very short time. In our target detection scenario, this feature vector is computed for every image chip, and we explore how well it may represent the overall information of the chip so as to support classification (e.g., chips containing boats might have significantly different gist signatures than chips which do not). Saliency and gist features appear to be complementary opposites [50]: saliency fea-

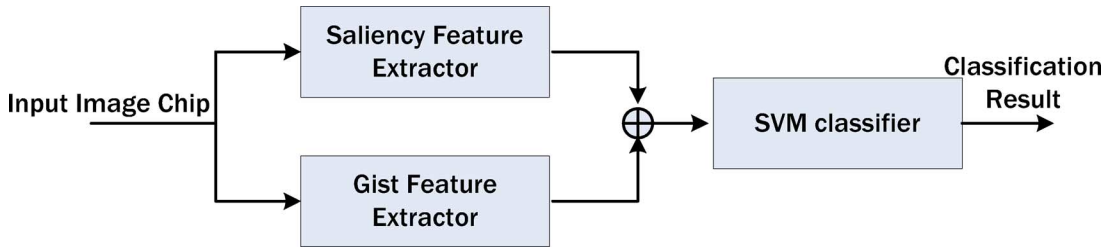



Fig. 1. Diagram of the image classification system applied to every image chip.

tures tend to capture and summarize the intensity and spatial distribution of those objects within a chip which stand out by being significantly different from their neighbors, while gist features capture and summarize the overall statistics and contextual information over the entire chip.

Given the proposed chip-based analysis approach, the task of answering "where is the target?" is equivalent to answering "does this image chip include the target?" for every chip in a large image. To achieve this decision making task, a Support Vector Machine (SVM) [51], [52] is adopted as the classifier, while the biologically inspired saliency-gist features are explored to form the feature vector in the feature space. The system overview diagram can be seen in Fig. 1.

## II. Design and Implementation

Here we first describe the two computational models proposed to compute the saliency features and gist features separately.

### A. Saliency Feature Computation

We compute saliency maps using several variants of the general Itti-Koch [31] architecture, and we then compute basic saliency map statistics for each variant. While in the original model only simple biological features (color, intensity, orientation) were employed, we here develop several new features which might be more effective in supporting the target/non-target classification task. The block diagram of the proposed model is shown in Fig. 2. In this model, an image is analyzed along multiple low-level feature channels to give rise to multiscale feature maps, which, as in the original Itti-Koch model, detect potentially interesting local spatial outliers. Ten feature channels are adopted in this paper: intensity, orientation (0°, 45°, 90° and 135°, combined into one "orientation" channel), local variance, entropy, spatial correlation, T-junctions, L-junctions, X-junctions, endpoints and surprise. Note that color information is not used since the images often are greyscale. Some of these feature channels (variance, entropy, spatial correlation) are computed by analyzing $16 \times 16$ image patches, giving rise to a map that is 16 times smaller than the original image horizontally and vertically (one map pixel per $16 \times 16$ image patch). The remaining feature channels are computed using image pyramids and center-surround differences, as in the original Itti-Koch algorithm: for each of these feature channels, center-surround scales are obtained from dyadic pyramids with nine scales, from scale 0 (the original image) to scale 8 (the image reduced by factor to $2^8 = 256$ in both the horizontal and vertical dimensions). Six center surround difference maps are then computed as point-to-point difference across pyramid scales, for combination of three center scales ($c = \{2,3,4\}$) and two center-surround scale differences ($\delta = \{3,4\}$). Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-scale competition for activity, followed by within-feature, across-scale competition. In this way, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. Feature maps then contribute additively to the corresponding saliency maps (SMs) that represent the conspicuity of each location in their channel. Finally, a saliency map feature extractor is applied to summarize each saliency map into a 4D vector with mean, variance, number of local maxima and average distance between locations of local maxima. All those feature vectors from the ten model variants are combined into a 40D vector referred to as the "saliency features." More information about the model is described in details in the following.

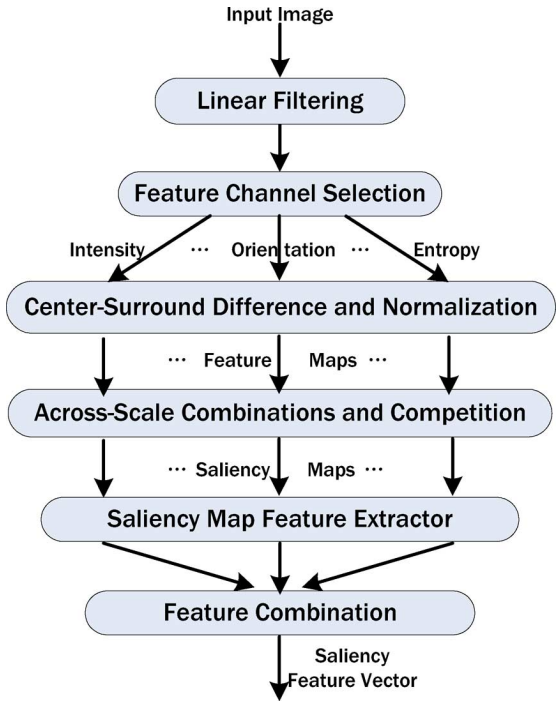



Fig. 2. Block diagram of the saliency features computation model applied to every image chip.

*Intensity Channel:* With the image chip as input, nine spatial scales are created using a dyadic Gaussian pyramid [25], which progressively low-pass filters and subsamples the input image, yielding horizontal and vertical image-resolution factors ranging from 1:1 (scale zeros) to 1:256 (scale nine).

Intensity represents the amount of light reflected by the corresponding point on the object in the direction of the camera view and multiplied by some constant factor that depends on the parameters of the imaging system. In our experiments, the range of the intensity value is from 0 to 65 535 (16-bit image) or from 0 to 255 (8-bit image) for all images $I_s$ $(s = 0, 1, \ldots, 8)$ at every spatial scale. This channel is essentially as previously described [25].

*Orientation Channel:* Orientation features are generally very effective feature in identifying objects, as demonstrated for example by humans' ability to understand line drawings. Here we adopt Gabor filters $(\theta_k = 0°, 45°, 90°, 135°)$ to extract the orientation feature. For each image I in the image pyramid, the orientation feature maps can be obtained as follows [25]:

$$M_{O,k} = \mathrm{Gabor}(I, \theta_k). \tag{1}$$

*Local Variance Channel:* Local variance channel is used to capture local pixel intensity variance over $16 \times 16$ image patches of the image chip of interest. This feature is of interest here as it has previously been shown to attract human attention [53], [54]. For each $16 \times 16$ image patch, the local variance feature map can be computed as follows:

$$M_V(i,j) = \mathrm{sqrt}\left(\frac{\sum_{sz} I^2(i,j) - S_{sz} * \mathrm{Mean}(I_{sz}(i,j))}{S_{sz} - 1}\right) \tag{2}$$

here $S_{sz}$ is the total pixel number of pixel $(i, j)$'s neighborhood with size of $sz$ ($\mathrm{sz} = 16 \times 16$ in our implementation).

*Entropy Channel:* Entropy as implemented here also provides a simple measure of information content in small $16 \times 16$ image patches. We follow the definition proposed by Privitera and Stark [54] who showed that such measure of entropy also correlates with human eye fixations. Note that many more sophisticated measures of entropy could be computed at the chip, image, or image sequence level, but this one has the advantage of being simple and motivated by previous human gaze tracking experiments. In image processing, entropy always indicates the probability distribution of the image intensity. The entropy value can be computed with the formula described in the following:

$$M_E(i,j) = -\sum_{I \in I_{sz}} p(I) * \log(p(I)) \tag{3}$$

where $I_{sz}$ means the neighborhood of the pixel at $(i, j)$ location, $p(I)$ stands for the probability of possible intensity $I$ in its neighborhood.

*Spatial Correlation Channel:* For two random variables X and Y, their correlation can be formulated as

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}. \tag{4}$$

Here, spatial correlation is computed at every location between a local $16 \times 16$ patch and other patches at a given radius from the local patch. It represents the similarity between the local patch and its neighbors. In the spatial correlation saliency map, low spatial correlation is a simple measure of high salience, i.e., low similarity.

*Junction Channels:* In addition to the Orientation channel described previously, several "junction" channels are created to further characterize the edge contents of image chips. Taking the local edge responses in different directions over small image patches into consideration, four different kinds of junction channels are created, all included in the junction saliency map: L-junction, T-junction, X-junction and endpoint. The L-junction channel is sensitive to "corner" features: it responds at locations where two edges meet perpendicularly and end at the intersection point. The T-junction channel responds when two edges are perpendicular and only one of them ends at the intersection point. Likewise, the difference of X-junction from T-junction is that in X-junction both edges do not end at the intersection point. Finally, the Endpoint channel responds when an extended edge ends. All junction channels are computed using a common framework which considers the collection of edge responses from the four maps in the Orientation channel, at points neighboring the point of interest.

We consider the 8-neighborhood of a given point of interest (at a given scale between 0 and 8 in our pyramid framework), and the one of the four orientation responses at each of the eight neighbors which is along the line segment from the central point to the neighbor (e.g., at the neighbor above the central point, the vertical orientation response is considered; at the neighbor to the left of the central point, the horizontal orientation response is considered). The response characteristics of a given junction channel is then given by a disjunction (sum) of binary response patterns (binary filter masks) applied to the neighbors' responses. For example, the T-junction detector will respond to 1) for an upright T, responses to the left (and from the orientation channel for horizontal orientation), right (horizontal orientation), and below (vertical orientation) the point of interest, plus 2) for a T rotated 90° clockwise, responses above, below, and to the left, plus 3) for an upside-down T, responses above, left and right, plus 4) for a T rotated 90° counter-clockwise, responses above, below and to the right. The L-junction and X-junction channels are defined likewise, and the mask pattern for the endpoint channel is simpler, as it will simply require that an orientation response exists on one side of the point of interest but not on the other (for example, some vertical response above but none below).

*Surprise Channel:* We recently proposed an enhanced saliency model, which exploits a new Bayesian definition of surprise to predict human perceptual salience in space and time [55]–[57]. Very briefly, surprise quantifies the difference between prior and posterior beliefs of an observer as new data is observed. If observing new data causes the observer to significantly reevaluate his/her/its beliefs about the world, that observation will cause high surprise. Surprise complements Shannon's definition of information by emphasizing the effect of data observations onto the internal subjective beliefs of an observer, while Shannon information objectively characterizes

the data itself (in terms of, e.g., how costly it would be to transmit from one point to another). Here, we use this new model as well, though we only consider the spatial domain since all images are static. Surprise is then computed for each $16 \times 16$ image patch by establishing prior beliefs from a large neighborhood of image patches, and computing the extent to which such beliefs are adjusted into posterior beliefs after information about the central patch of interest is observed. The surprise map computed under these conditions is similar to a regular saliency map, except that the Bayesian surprise computations are used for competition across space instead of the mechanism described in the following. The surprise map is, thus, an optimized weighted combination of intensity, orientation and junction features, to which a spatial surprise detector is applied.

*Feature Maps Competition:* In all maps except surprise (which has its own internal competition dynamics), a feature map competition mechanism tends to globally promote maps in which a small number of strong peaks of conspicuous locations is present, while globally suppressing maps which contain numerous comparable peak responses. To implement this, first normalize the feature map to a fixed range $[0 \ldots M]$, and then find the global maximum value $M$ and the average value $\bar{m}$ of other local maximums, finally globally multiplying the map by $(M - \bar{m})^2$, as was previously described in detail [25].

*Saliency Map Feature Extractor:* For each of the ten variants of the model, the obtained saliency map is relatively high-dimensional data (for example, a $512 \times 512$ image chip's saliency map size is $32 \times 32 = 1024$D), and this becomes especially true when all ten channels' saliency maps are combined. To reduce the data dimensionality while keeping the most important information, we compute four summary statistic values to represent each saliency map: mean value $m_k$, standard deviation $v_k$ over the saliency map's pixels, number $n_k$ of local maxima (peaks) in the map, and the average Euclidean distance between the local maximum points $d_k$. The computation formulas are described as

$$m_k = \frac{1}{W \times H} \sum_{i,j} SM_k(i,j) \tag{5}$$

$$v_k = \sqrt{\frac{1}{Sz - 1} \sum_{i,j} (SM_k(i,j) - v_k)^2} \tag{6}$$

$$d_k = \text{mean}(\sqrt{(i_p - i_q)^2 + (j_p - j_q)^2)}) \quad p, q < n_k, p \neq q \tag{7}$$

where $W$ and $H$ are the saliency map size, and $Sz$ is the saliency map's area, $(i_p, j_p)$ and $(i_q, j_q)$ are local maximum points in saliency map, subscript $k$ indicates the saliency map type (intensity, orientation, ...). A rational explanation of this is that the saliency map describes the conspicuity of the image and only the most salient points or regions will show on the saliency map, therefore, we can use these four values to represent the most important information of the saliency map. We may lose the salient objects' position information, however, we hypothesize that it might not affect the performance of the detection task greatly: the four statistics should capture some information about the distribution of salient objects in the image chip, no matter where they are, and may serve as a useful position-invariant (and somewhat rotation- and scale-invariant) descriptor of the image chip. Our experiments shown in the following will directly test this hypothesis. According to the previously shown analysis, the dimension of the combined saliency feature vector is: $\text{Dim}_{\text{sal}} = N_{\text{feature Channels}} \times 4 = 10 \times 4 = 40$.

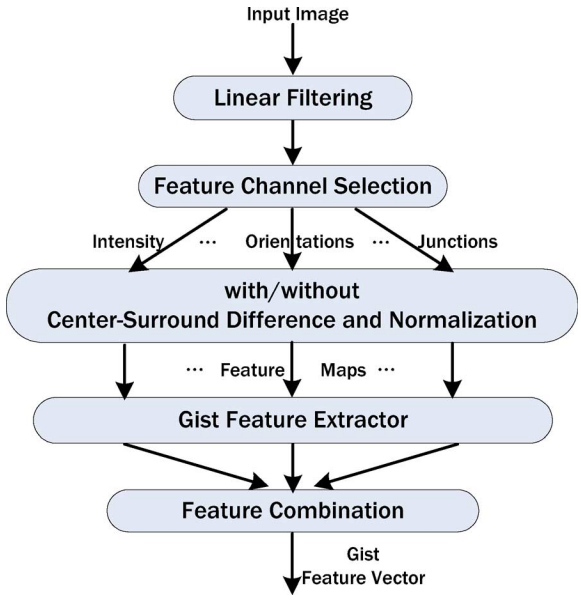



Fig. 3. Block diagram of gist features computation model applied to every image chip.

### B. Gist Feature Computation

The gist feature computation model [50] is related to the saliency computation model, except that it embodies concepts of feature cooperation across space rather than competition. The gist computation model architecture used in the present paper is shown in Fig. 3 and the low-level features channels include intensity, four orientations (0°, 45°, 90°, and 135°), and four L-junctions (0°, 45°, 90°, and 135°), four T-junctions (0°, 45°, 90°, and 135°), four endpoints (0°, 45°, 90°, and 135°) and X-junction, therefore, 18 different feature channels are adopted.

Unlike the saliency feature extraction model, both center-surround and raw (before center-surround) pyramid levels are exploited. For the center-surround operation, six center surround difference maps are then computed within each pyramid as point-to-point difference across pyramid scales, for combination of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$). For the raw operation, the adopted raw pyramid scales range from 0 to 4.

Since gist features describe an image chip's overall information, we only use mean value to represent each of the gist feature maps

$$G_{k,s,c} = \frac{1}{W \times H} \sum_{i,j} GF_{k,s,c}(i,j) \tag{8}$$

where $W$ and $H$ are the gist feature map size, indices k, s, c denote feature map type, scale, center-surround type,

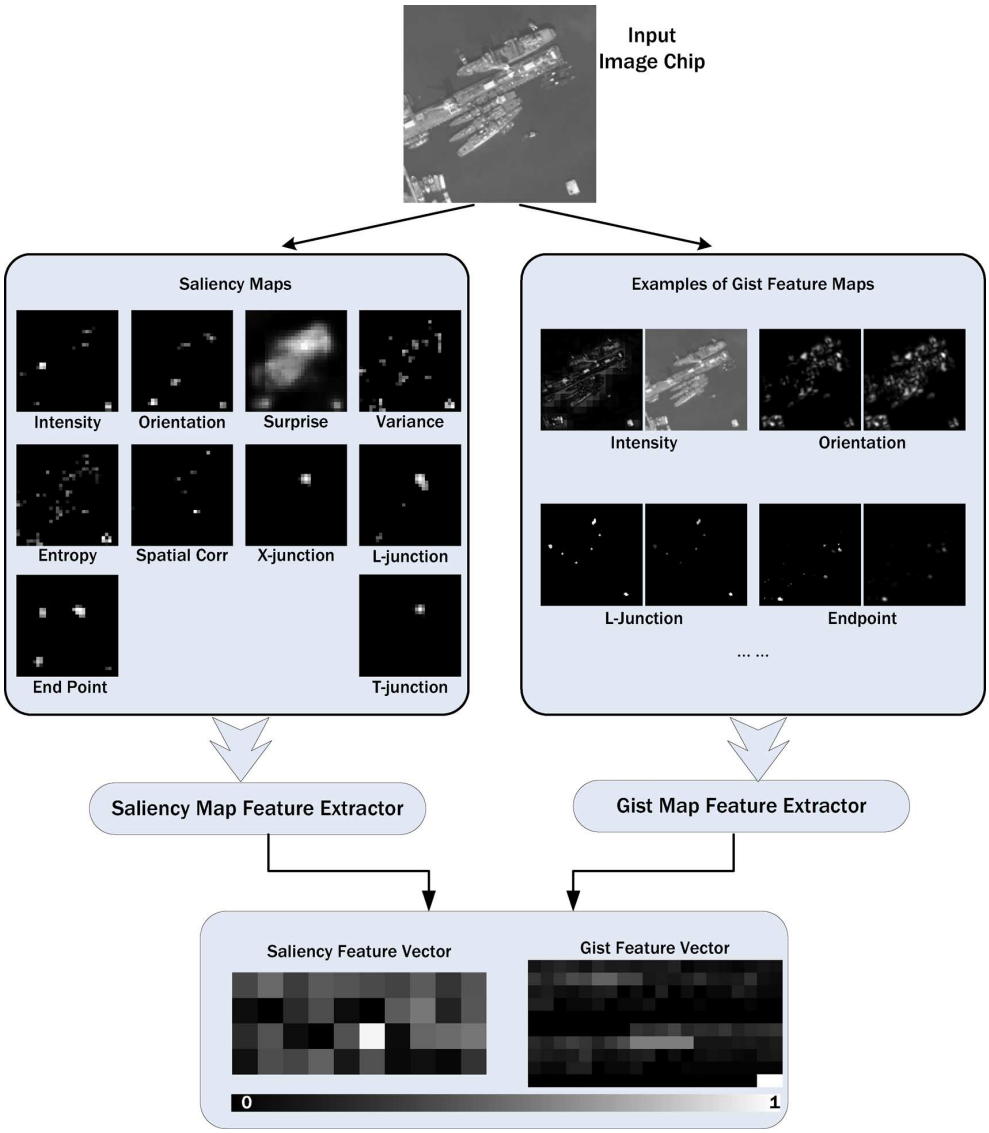



Fig. 4. Example of complete saliency-gist feature extraction for an image chip. Note that the saliency maps shown already have been subjected to spatial competition; hence, for example, out of the initially many responses in the T-junction channel at various locations and for various spatial scales, one ends up winning the competition strongly and dominating the other ones in the particular example image chip shown. The gist feature map examples shown in this figure are presented in pairs, for the center-surround and raw no-center-surround computations. In each pair, the left map is the center-surround result while the right map is the no-center-surround result. There are four pairs shown in this figure: intensity, 45° orientation, 135° L-junction and endpoint. The scales of center surround are 2 (center) and 5 (surround) while the scale of no center-surround is 2.

respectively. Therefore, the gist feature vector dimensions are $\mathrm{Dim}_{\mathrm{gist}} = N_{\mathrm{feature\ Channels}} \times (N_{\mathrm{Center\ Scales}} \times N_{\mathrm{Surround\ Scales}} + N_{No\ CS\ \mathrm{Scales}}) = 18 \times (3 \times 2 + 5) = 198$.

We simply combine the saliency features and gist features together to form the final saliency-gist feature vector, which is a $40 + 198 = 238$ dimensions vector. One example of the complete process for one input image is illustrated in Fig. 4. Before using these feature vectors to detect targets, it is necessary to normalize the feature values alone feature types. The normalized feature then can be sent to the classifier to implement detection task. Considering the high nonlinearity of the feature vectors' distribution, RBF (radial basic function) based SVM were adopted to complete the classification task. In this paper, SVM provided by [52] were adopted for its easy to use. Furthermore, for the normalized input, the parameters of SVM can be optimized automatically and no tuning is needed.

## III. Experiments and Results

We test the proposed model with four experiments of challenging broad area search in satellite images. Mainly, the search tasks are challenging because of high intraclass variability in the target category: boats in experiments 1 and 2 (from small vessels to large ships), buildings in experiment 3, and airplanes in experiment 4. To compare our algorithm to the state of the art, we decided to employ the HMAX [14], [18], SIFT [7], and the hidden scale salient structure object detection algorithm [16] as references. We opted for HMAX and SIFT because of their popularity in target detection and in generalization over object cate-

Fig. 5. Examples of target image chip and no target image chip for experiment 1, detecting image chips which contain one or more boat(s) of any size and type. The top-row image chips include one or more target boat, while the bottom-row images do not include any target.

gories from limited training data. The hidden scale salient structure method is similar to our research and performs very well in target detection for satellite images. All these references' source code is available and, thus, easy to implement for our experiments. To complement our analysis, we also compare our algorithm to Siagian-Itti's gist features proposed in [50] to show how much is gained from our very simple 4D summaries of saliency maps and from the new gist features used here.

### A. Experiment 1

The first dataset (dataset 1) used to test the proposed model includes 14 416 image chips ($500 \times 500$) which were cut out of one large broad-area satellite image (size $21\,500 \times 27\,500$) with a slide window step size of 200 pixels (hence, two successive chips overlap by 300 pixels). All target centers in the broad-area image were manually labeled as ground truth (if several boats were connected together, then we treated them as one target); the boats' sizes ranged from tens of pixels to hundreds of pixels. Among these image chips, 705 included targets (various boats). Examples of target image chips and nontarget image chips can be seen in Fig. 5. To compare the effectiveness of the proposed saliency-gist approach to the state of the art, we compare it with the gist feature proposed in [50] (here we call it standard gist feature), the HMAX feature [14], [18], the SIFT feature [7] and the hidden scale salient structure feature [16].

In the classification step, N positive image chips (which include one or more targets) and N negative image chips (which do not include any target) are randomly selected from the dataset and used as the training samples, while all remaining image chips are treated as test data. The commonly used measurement to evaluate the precision of classification are percentage of true positive (TP) and true negative (TN) which are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{9}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\% \tag{10}$$

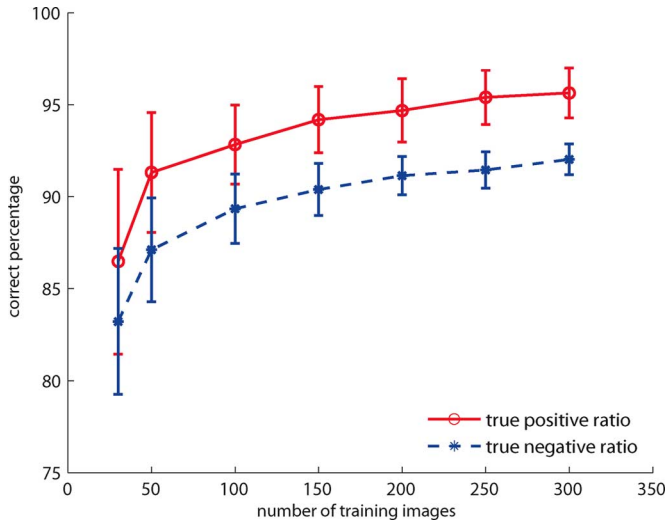



Fig. 6. Classification results for experiment 1 (detecting boats), for different numbers of training images from the pool of 705 total available chips containing one or more targets (error bars are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

where TPR and TNR stands for true positive ratio and true negative ratio. The classification results with different numbers of training samples are shown in Fig. 6. It is easy to see that when we increase the number of training samples, the classification rate improves. It is worth noting how, even with a small number of training samples, the results do not catastrophically degrade but rather remain quite high (above 80% hits and correct rejections).

For a classification system, pursuing higher TPR and lower false positive ratio (FPR) usually contradict each other: a higher TPR often causes higher FPR. With different decision criteria, the classification results may vary. For example, in a warning system, pursuing higher TPR is preferred to pursuing lower FPR. Since the receiver operating characteristic (ROC) curve has the ability to show the comparison of TPR and FPR as the

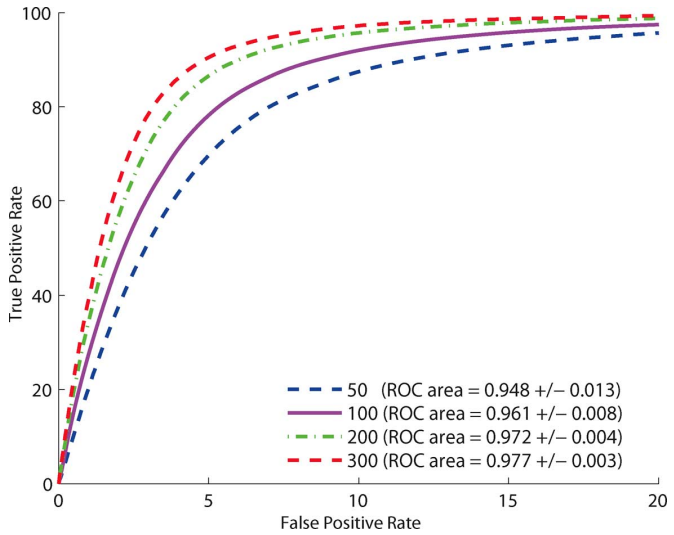



Fig. 7. ROC curve for the proposed system (zoomed-in on the horizontal axis) for different numbers of training samples, for experiment 1 (detecting boats). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

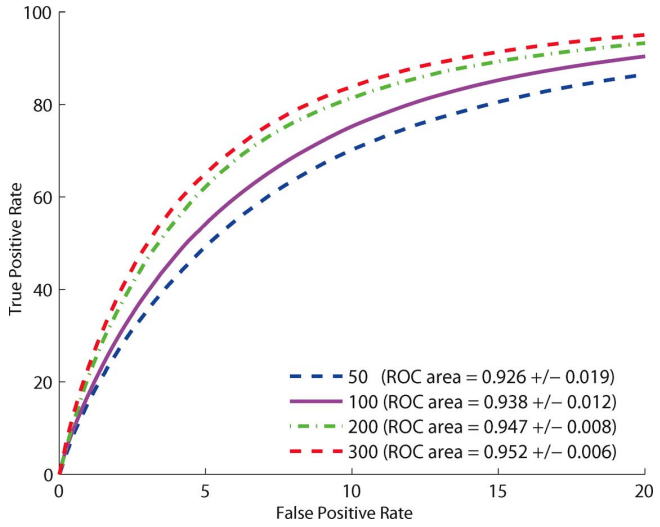



Fig. 9. ROC curve (zoomed-in on the horizontal axis) for different numbers of training samples, for experiment 2 (detecting boats, with training set from experiment 1). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

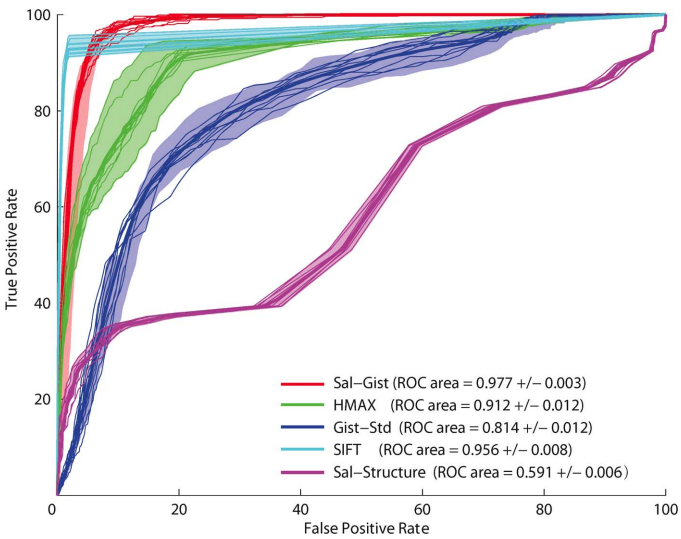



Fig. 8. ROC curve comparison among different feature types in experiment 1, detecting boats. 300 training samples were used for both the positive and negative target categories. The mean ROC area values (corresponding to the thick curves) and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for saliency-gist feature, standard gist feature, SIFT feature and hidden scale salient structure feature, and from a smaller number of ten runs for HMAX feature because of the high run-time of HMAX). The shadow envelopes and ten thin curves for each model show the ROC curves which reach the maximum and minimum ROC area in the multiple runs of the experiment (using different randomly-chosen training samples from the training set). ROC performance for the proposed Sal-Gist algorithm is significantly better than for all other methods.

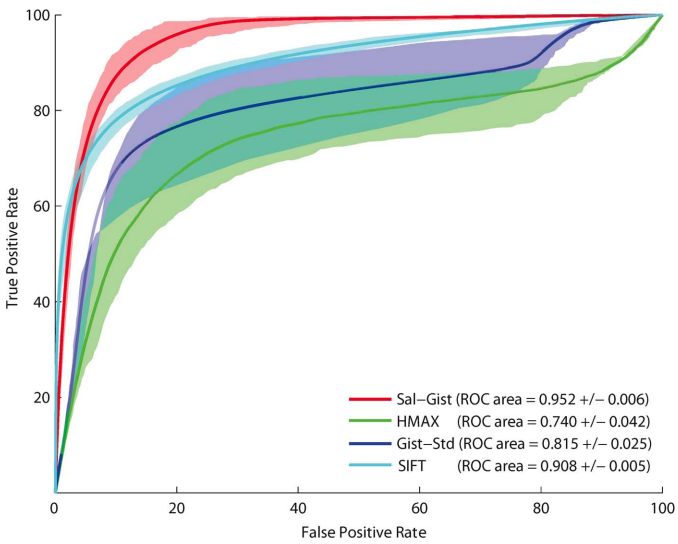



Fig. 10. ROC curve comparison among different feature types in experiment 2, detecting boats. 300 training samples were used for both the positive and negative target categories (from experiment 1's dataset). The corresponding mean ROC area values and standard deviations are labeled in the legend (100 runs for saliency-gist feature, standard gist feature and SIFT feature, ten runs for HMAX feature because of the high complexity). The shadow contours stand for the ROC curves which reach the maximum and minimum ROC area in multiple experiment runs.

classification decision criterion changes, it is widely adopted to compare performance of two different classification systems. A higher TPR while low FPR stands for a better classification system, and usually this can be described by the area under the ROC curve. An ROC area equals to 1 means a system that can perfectly classify the categories without any error, an ROC area equals to 0.5 stands for a random classification system, and, the bigger ROC area, the better classification performance. To compute ROC curves with our algorithm, we systematically vary distance to the decision boundary as the criterion parameter. Fig. 7 shows ROC curves for the proposed saliency-gist algorithm, as a function of the number of training examples. We can see that performance degrades gracefully as the number of training examples is decreased. The corresponding ROC curves and the ROC areas of classification with saliency-gist feature,

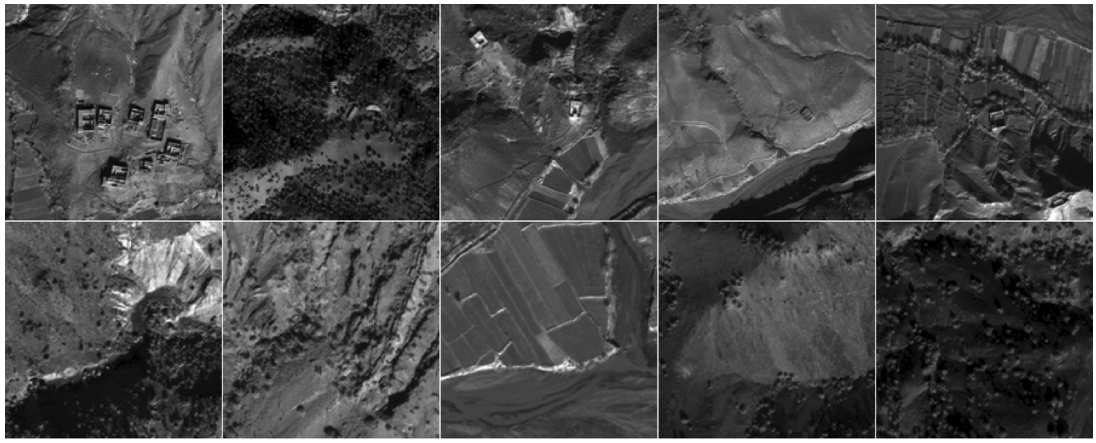


Fig. 11. Examples of target image and no target image for experiment 3, finding buildings of any type, size, and style. Top-row images include one or more target(s) while the bottom row images do not include any target.

HMAX feature, SIFT feature, hidden scale salient structure feature and standard gist features are shown in Fig. 8. (marked as sal-gist, HMAX, SIFT, sal-structure, and gist-std in the figure, respectively). It is clear from the figure that the saliency-gist feature outperforms the other features greatly (t-tests on the 100 ROC values obtained with each of the 100 randomly selected training sets, $p < 10^{-10}$ or better), hence, demonstrating appeal of the proposed approach. Also, from the figure we can see that the hidden scale salient structure method almost failed in this experiment. This is mainly because the targets (boats) are not salient compared with many inland buildings when using the salient structure algorithm in [16] and, thus, the algorithm misclassified many buildings as boat targets.

### B. Experiment 2

This experiment tests how training on one broad-area image taken at one given time and location may generalize to testing on another broad-area image taken at another time and location. The second dataset (dataset 2) includes 11 385 image chips ($500 \times 500$) which were cut out of another large broad-area satellite image (size 23 $300 \times 20$ 100, taken from the same country but on a different date and at a different place than the broad-area image of experiment 1), with the same slide window size as in dataset 1. We labeled the targets manually as ground truth like in experiment 1 and there are 1 049 image chips which include one or more target(s). In this experiment, training samples for the classifier are randomly selected from dataset 1, while all the image chips in dataset 2 are used as test set.

Fig. 9 shows that ROC performance improves with the number of training samples, as in experiment 1. With 300 training samples, ROC area was 0.952 here, as compared to 0.977 in experiment 1 (Fig. 8), suggesting good generalization capability to new, never seen images. Like in experiment 1, the comparison of detection results with the saliency-gist feature, standard gist features, HMAX features, and SIFT features (the hidden scale salient structure feature is not adopted to do the comparison in this experiment due to its poor performance in experiment 1) shows that the saliency-gist feature performs much better than other three features (Fig. 10, t-tests, $p < 10^{-12}$ or better).

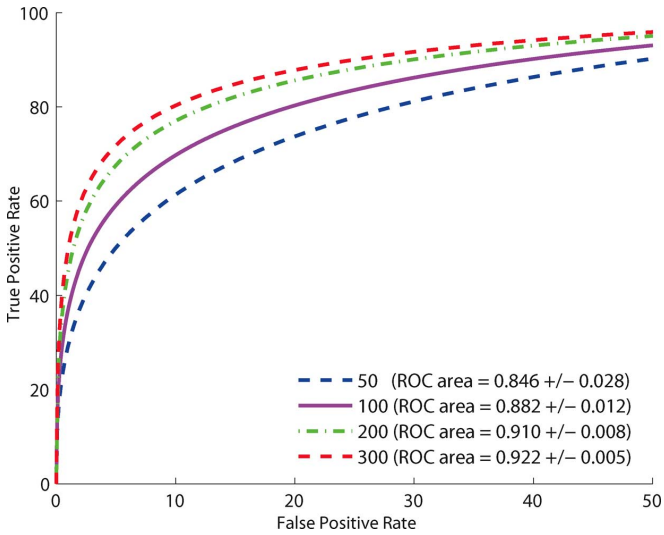



Fig. 12. ROC cures (zoomed in on the horizontal axis) for different numbers of training samples in experiment 3 (detecting buildings). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

### C. Experiment 3

In this experiment, targets are simply defined as "buildings" in satellite images. This experiment, thus, tests the ability of our same algorithm to classify very different types of targets; the intraclass variability here is also arguably even larger than in experiments 1 and 2 (see Fig. 11). The dataset (dataset 3) used here includes 108 885 image chips (this experiment used a smaller chip size of $256 \times 256$ because the targets were also smaller than in experiments 1 and 2) with 6 323 of them being positive examples. Fig. 11 shows examples of buildings and negative examples. Like in experiments 1 and 2, the image chips were cut from a broad-area satellite image (size 16 $512 \times 27$ 520, taken from a different country and a different year than the images of experiments 1 and 2). The slide window size here was 64 pixels. Ground-truth information for this dataset (locations of buildings) was provided to us by an outside corporation. The ROC curves for different numbers of training samples are plotted in Fig. 12. As we can see, performance again improves

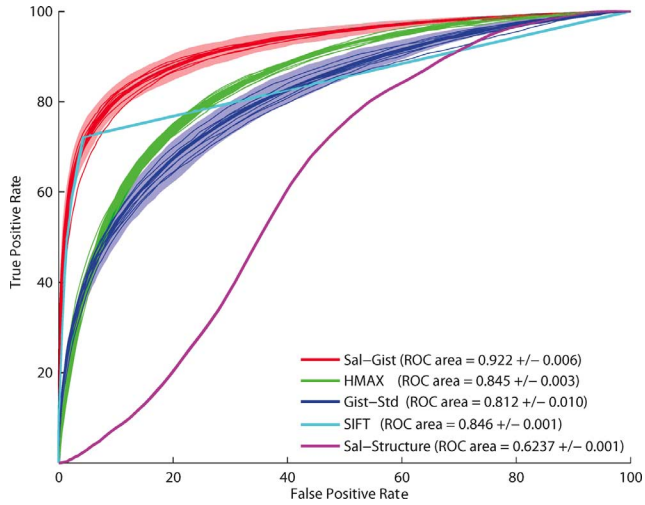



Fig. 13. ROC curve comparison among different feature types in experiment 3, detecting buildings. 300 positive and 300 negative training examples were used. The experiment parameters are the same as in experiments 1 and 2. The shadow contours of SIFT and hidden scale salient structure feature are quite small and can not seen in this figure.

with the size of the training set. Again, we compare the detection results with the standard gist features, the HMAX features, the SIFT features and the hidden scale salient structure features. The corresponding ROC curves of classification with these feature types are shown in Fig. 13. It is clear from the figure that the saliency-gist features again outperform the other two features greatly (t-tests, $p < 10^{-17}$ or better).

To illustrate the detection result in a more straightforward and global way, we adopt a probability map representation (PM) to show the results. A probability map is a matrix which depicts the probability value for each image chip to contain a target. The rescaled broad-area satellite image and some example target buildings are shown in Fig. 14(a), and the corresponding probability map is shown in Fig. 14(b), the red points in the images stand for the labeled targets' center location. This simple representation reinforces the ROC results and suggests a high performance of the algorithm, as shown by the overlap between red ground truth locations and brighter locations in the PM (higher probability of target according to our algorithm). During search for buildings, exploring the image in decreasing order of target probability per our algorithm would isolate more targets faster than a naïve scan from left to right and top to bottom.

### D. Experiment 4

An aerial image of an airport is adopted in this experiment to detect the "airplanes" (see Fig. 15). The dataset (dataset 4) used here include 2 601 image chips, of which 1 382 of them include a target. For each chip, the size is $64 \times 64$ due to the small target size. Compared to the previous experiments, the target is relatively easier to detect because intra class variances (both in shape and area) are small. Here we compared the detection performance among saliency-gist feature, hidden scale salient structure feature and SIFT. Ten positive and ten negative examples were randomly selected as training data while the rest were taken as test data. The detection results from different methods are plotted in Fig. 16 (100 runs for each). From the figure we can see that all three methods perform very well while the proposed method performs even better than the others (no shadow contours plotted here because the difference of results from different method is small while the variance of result from SIFT is relatively big which may cause the whole figure not clear).

### E. Saliency Versus Gist

As saliency-gist features yield great classification results, it is interesting to see the separate contributions of the saliency features and gist features. The ROC area of using saliency features only, gist features only (in our new implementation, which includes more feature channels than the older Gist-Std model), and combined saliency-gist features in all four experiments are shown in Table I. It can be seen from the table that the combined saliency-gist features outperforms both saliency features and gist features in all experiments. Hence, these results show that the saliency features and gist features are not fully redundant, even though they are computed using similar low-level feature detectors. In addition, the table shows that saliency features perform better in experiments 1, 2 and 4, while gist features perform better in experiment 3. Thus, in different cases, the classification results depends more on different types of information (saliency information and gist information), which again reinforces the benefits of using both types of features.

## IV. Discussion

Our results show that the proposed algorithm performs better than the state-of-the-art (HMAX algorithm, SIFT algorithm, hidden scale salient structure algorithm and previously proposed gist algorithm alone) in difficult target search scenarios. This was achieved in situations where targets can vary greatly in their size, shape, and number of targets per chip. Overall, the proposed algorithm is conceptually very simple and at the same time very general, since the feature extraction stages were not designed or tuned for the specific types of images and targets tested here. Taking all results together suggests that the proposed system may be further applicable to a wide range of images and target types. Indeed, nothing in the proposed algorithm has been specifically developed or tuned for the boat or building or airplane targets tested here, or for the type of images processed in our experiments.

The success of the proposed approach may be due to our use of two complementary sets of biologically-inspired features: gist features largely discard spatial information, while saliency features summarize it. In the human brain, it is clear that object recognition relies on being able to compute invariants, but at the same time pose parameters are not lost: although one recognizes an upside-down face as being a face, one is also aware that it is upside-down. Our approach here seems to benefit from this dual view of the image data. Recently, some other biologically-inspired feature extraction methods [19] have started to use the "gestalt" information (continuity, symmetry, closure, repetition, etc.) to conduct object detection and have shown promising results. It is likely that combining these feature types will get even better detection performance. There are many other feature types which could be also added to our approach, including for example locally-binary pattern (LBP) features which have been particularly successful in texture segmentation [58].

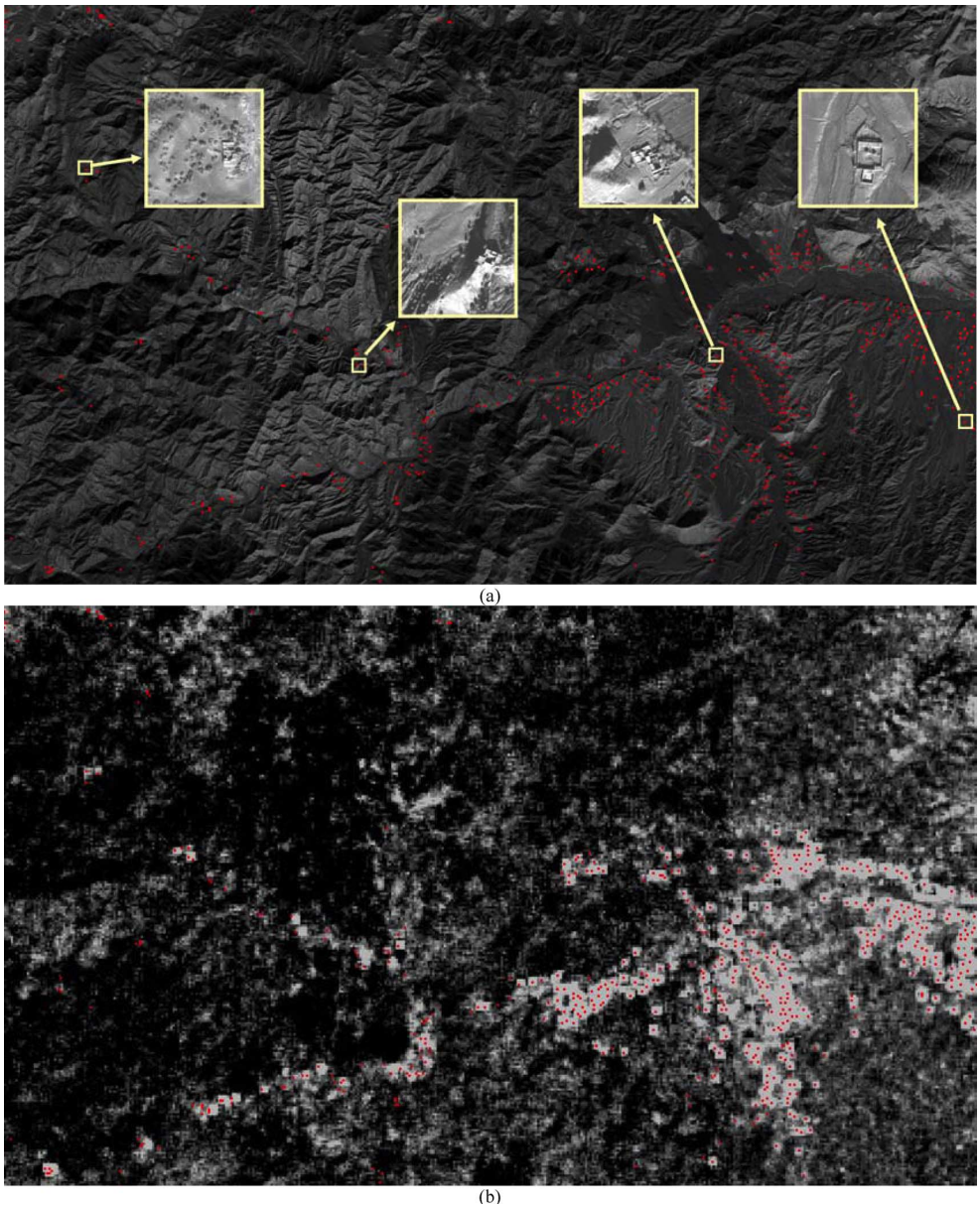


Fig. 14. Illustration of "building" detection in experiment 3. (a) Rescaled broad-area satellite image (16 512 × 27 520 pixels) and some target examples. (b) Probability map of (a) computed by our algorithm. The red points are the true target center locations. In the probability map, lighter areas indicate higher probability of targets, while darker areas denote lower probability of targets according to the algorithm.

TABLE I
COMPARISON OF ROC AREAS OF DIFFERENT TYPES OF FEATURES IN FOUR EXPERIMENTS

| | Saliency Feature | Gist Feature | Saliency-Gist Feature |
|---|---|---|---|
| Experiment 1 | 0.969 ± 0.003 | 0.943 ± 0.008 | 0.977 ± 0.003 |
| Experiment 2 | 0.945 ± 0.007 | 0.903 ± 0.009 | 0.952 ± 0.005 |
| Experiment 3 | 0.789 ± 0.007 | 0.905 ± 0.005 | 0.922 ± 0.005 |
| Experiment 4 | 0.927 ± 0.031 | 0.942 ± 0.028 | 0.976 ± 0.003 |

The proposed algorithm does not take any complex procedure to combine the features extracted, although many research studies have proposed feature combination algorithms to improve classification performance [59], [60]. Here we only show that the combination of gist feature and salient feature are complementary and can achieve good performance in target detection. It is interesting that saliency and gist features both contribute significantly to performance, and are not fully redundant (Table I). This suggests a new use of saliency algorithms, for classification of images based on their saliency maps, as opposed to using the saliency maps to generate shifts of attention. It is interesting to think whether humans and other animals may use this as well. It is possible that human saliency maps in posterior parietal cortex, the pulvinar nucleus, the frontal eye fields,

Fig. 15. Image used to detect the airplanes in experiment 4.

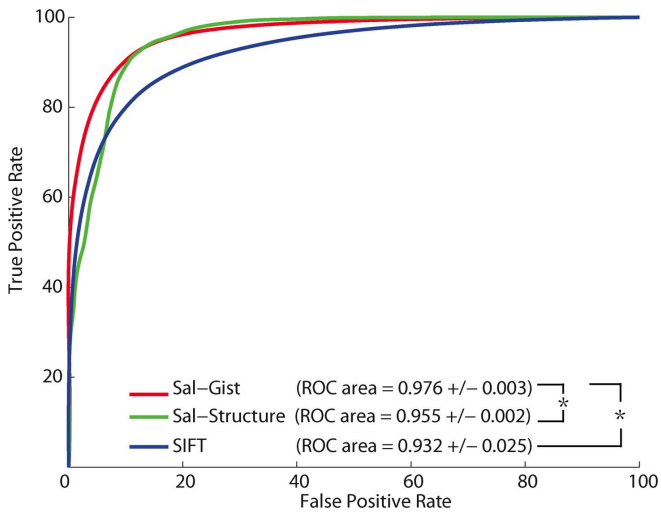



Fig. 16. ROC curve comparison among different feature types in experiment 4, detecting airplanes. Ten positive and ten negative training examples were used. * indicate statistically different ROC performance (t-test, $p < 10^{-10}$ or better).

or the superior colliculus [31] may also be analyzed in a holistic fashion and may contribute to the very rapid understanding of the rough layout of the scene. That is, the coarse structure of saliency maps may combine with the broad semantic information provided by the gist features to yield a coarse and rapid understanding of both a scene's gist and layout [61].

Our approach reinforces the idea, as shown by recent successes in the domains of statistical machine translation of text into foreign languages or of speech analysis [62], that relatively shallow statistical analysis of large datasets can yield surprisingly good classification and recognition results. Indeed, our algorithm does not try to understand the geometric structure or other specific high-level or cognitive feature of targets (e.g., buildings should have walls, tend to be rectangular, etc) and is not attempting recognition by components (breaking down target objects into elementary parts and their spatial arrangements [63].

The proposed algorithm is mostly intended as a front-end, to be used to perform coarse preliminary analysis of large complex scenes. The data returned certainly is still far from representing a complete understanding of the scene's contents. However, our algorithm's output can be used in at least two practical ways: first, to compute statistics at the region level, like, e.g., finding areas in the world with high concentrations of boats, or determining which regions in a country have more buildings and, hence, may be more densely populated. Such basic statistics may be of great use on their own, for example when planning rescue efforts following a natural disaster, or may assist a human image analyst in performing deeper and more cognitively-driven surveys of imagery. Second, our algorithm can be used to rank image chips by interest (using the probability maps of Fig. 14), so as to focus limited resources onto the most promising image locations. Resources may be limited because of limited human personnel, human viewing time (e.g., when using rapid serial visual presentation of image chips [64], or computation time (e.g., using a more sophisticated and time-consuming object recognition back-end to validate high-probability chips). It is likely that our system could perform even better if one was to apply some of the recognition-by-components principles or other recognition back-end to the high-probability target chips returned by our algorithm.

Thus far, our algorithm has only been applied to greyscale visible imagery. With the increasing popularity of color and multispectral imagery, it remains to be tested in future work whether our simple approach will scale up to a larger number of spectral bands. All C++ source code for our algorithms is available on the authors' web site (http://iLab.usc.edu).

## ACKNOWLEDGMENT

The authors would like to thank DigiGlobe for providing the aerial images. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States Government or any agency thereof.

## REFERENCES

[1] Y. Amit, *2D Object Detection and Recognition, Models, Algorithms and Networks*. Cambridge, MA: MIT Press, 2002.
[2] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2201–2216, Nov. 2008.
[3] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.
[4] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Hoboken, NJ: Wiley, 2009.
[5] I. Craw, H. Ellis, and J. Lishman, "Automatic extraction of face features," *Pattern Recognit. Lett.*, vol. 5, pp. 183–187, 1987.
[6] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, 1992.
[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[8] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.
[9] K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," in *Proc. ICCV*, 2005, vol. 2, pp. 1458–1465.
[10] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416–431, Mar. 2006.
[11] G. Bouchard and B. Triggs, "Hierarchical part-based visual object categorization," in *Proc. CVPR*, 2005, vol. 1, pp. 710–715.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale invariant learning," in *Proc. CVPR*, 2003, vol. 2, pp. 264–271.
[13] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondence," in *Proc. CVPR*, 2005, vol. 1, pp. 26–33.
[14] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, pp. 1019–1025, 1999.
[15] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. CVPR*, 2005, vol. 2, pp. 994–1000.
[16] B. Chalmond, B. Francesconi, and S. Herbin, "Using hidden scale for salient object detection," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2644–1655, Sep. 2006.
[17] J. Mutch and D. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," *Int. J. Comput. Vis.*, vol. 80, no. 1, pp. 45–57, 2008.
[18] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.
[19] S. Bileschi and L. Wolf, "Image representations beyond histograms of gradients: The role of Gestalt descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007.
[20] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Lab. J.*, vol. 14, no. 1, 2003.
[21] C. Chang, H. Ren, and S. Chiang, "Real-time processing algorithms for target detection and classification in hyperstpectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 4, pp. 760–768, Apr. 2001.
[22] H. Li and J. H. Michels, "Parametric adaptive signal detection for hyperspectral imaging," *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2704–2715, Jul. 2006.
[23] J. Lanir and M. Maltz, "Analyzing target detection performance with multispectral fused images," in *Proc. SPIE*, 2006.
[24] S. Buganim and S. R. Rotman, "Matched filters for multispectral point target detection," in *Proc. SPIE*, 2006.
[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[26] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, no. 3:3, pp. 1–15, 2008.
[27] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.
[28] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cogn.*, vol. 12, pp. 1093–1123, 2005.
[29] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–137, 1980.
[30] J. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonom. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.
[31] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.
[32] N. Bruce and J. Tsotsos, "Saliency, attention and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, 2009.
[33] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proc. IEEE CVPR*, 2005.
[34] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in highly dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.
[35] [Online]. Available: http://ilab.usc.edu/toolkit
[36] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Rev. Neurosci.*, vol. 5, pp. 495–501, 2004.
[37] M. W. Cannon and S. C. Fullenkamp, "Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations," *Vis. Res.*, vol. 31, pp. 1985–1998, 1991.
[38] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, "Visual cortical mechanisms detecting focal orientation discontinuities," *Nature*, vol. 378, pp. 492–496, 1995.
[39] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vis. Res.*, vol. 46, no. 26, pp. 4333–4345, 2006.
[40] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10–12, pp. 1489–1506, 2000.
[41] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[42] M. C. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, 1975.
[43] I. Biederman, "Do background depth gradients facilitate object identification?," *Perception*, vol. 10, pp. 573–578, 1982.
[44] B. Tversky and K. Hemenway, "Categories of the environmental scenes," *Cogn. Psychol.*, vol. 15, pp. 121–149, 1983.
[45] A. Oliva and P. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," *Cogn. Psychol.*, vol. 34, pp. 72–107, 1997.
[46] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, 2000.
[47] T. Sanocki and W. Epstein, "Priming spatial layout of scenes," *Psychol. Sci.*, vol. 8, pp. 374–378, 1997.
[48] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1995.
[49] R. Epstein, D. Stanley, A. Harris, and N. Kanwisher, "The parahippocampal place area: Perception, encoding, or memory retrieval?," *Neuron*, vol. 23, pp. 115–125, 2000.
[50] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.
[51] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE CVPR*, 1997, vol. 1, pp. 130–136.
[52] [Online]. Available: http://www.kernel-machines.org
[53] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Comput. Neural Syst.*, vol. 10, pp. 341–350, 1999.
[54] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison witheye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.
[55] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE CVPR*, 2005, vol. 1, pp. 631–637.
[56] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 547–554, 2006.
[57] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.
[58] T. Ojala and M. Pietikäinen, "Unsupervised texture segmentation using feature distributions," *Pattern Recognit.*, vol. 32, pp. 477–486, 1999.
[59] L. Wolf, S. Bileschi, and E. Meyers, "Perception strategies in hierarchical vision systems," in *Proc. IEEE CVPR*, 2006.
[60] I. Oh, J. Lee, and C. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 1089–1094, Oct. 1999.
[61] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, 2000.
[62] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Ling.*, vol. 29, no. 1, pp. 19–51, 2003.
[63] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychol. Rev.*, vol. 94, no. 2, pp. 115–147, 1987.
[64] W. Einhaeuser, T. N. Mundhenk, P. F. Baldi, C. Koch, and L. Itti, "A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition," *J. Vis.*, vol. 7, no. 10, pp. 1–13, Jul. 2007.

**Zhicheng Li** received the B.S. degree in electronics and information from Northwestern Polytechnical University, Xi'an, China, in 2005, and is currently pursuing the Ph.D. degree in school of automation science and electrical engineering, Beihang University, Beijing, China.

He is a Visiting Scholar in the Computer Science Department, University of Southern California, Los Angeles, since 2007. His main research interests include visual attention modeling, video compression, and target detection.

**Laurent Itti** received the M.S. degree in image processing from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1994, and the Ph.D. degree in computation and neural systems from the California Institute of Technology (Caltech), Pasadena, in 2000.

He is currently an Associate Professor of Computer Science, Psychology, and Neuroscience at the University of Southern California, Los Angeles. His research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, gist, saliency, and surprise, with applications to video compression, target detection, and robotics.

J. Neural Eng. **7** (2010) 016006 (10pp) doi:10.1088/1741-2560/7/1/016006

# Saliency-based image processing for retinal prostheses

**N Parikh**[1], **L Itti**[2] **and J Weiland**[3]

[1] Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA
[2] Department of Computer Science, Psychology and Neuroscience, University of Southern California, Los Angeles, CA, USA
[3] Departments of Ophthalmology and Biomedical Engineering, University of Southern California, Los Angeles, CA, USA

E-mail: njparikh@usc.edu, itti@usc.edu and jweiland@usc.edu

Received 31 August 2009
Accepted for publication 7 December 2009
Published 14 January 2010
Online at stacks.iop.org/JNE/7/016006

**Abstract**

We present a computationally efficient model for detecting salient regions in an image frame. The model when implemented on a portable, wearable system can be used in conjunction with a retinal prosthesis, to identify important objects that a retinal prosthesis patient may not be able to see due to implant limitations. The model is based on an earlier saliency detection model but has a reduced number of parallel streams. Results of a comparison between the areas detected as salient by the algorithm and areas gazed at by human subjects in a set of images show a correspondence which is greater than what would be expected by chance. Initial results for a comparison of the execution speed of the two algorithm models for each frame on the TMS320 DM642 Texas Instruments Digital Signal Processor suggest that the proposed model is approximately ten times faster than the original saliency model.

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

An electronic retinal prosthesis is under development, to treat blinding diseases like retinitis pigmentosa (RP) and age-related macular degeneration (AMD) [1]. In RP and AMD, the photoreceptor cells are affected while other retinal cells remain relatively intact. Photoreceptor cells convert light information entering the retina into electrical signals and hence the progressive loss of these cells leads to a gradual loss of vision in patients. The retinal prosthesis aims to provide partial vision by electrically activating the remaining cells of the retina. Current retinal prosthesis prototypes use external components to acquire and code image data for transmission to an implanted retinal stimulator. The external system consists of a small camera to capture video in real time and a portable video processor to convert image data to a series of command signals which are wirelessly transmitted to the implanted retinal stimulator.

Human monocular vision has a field of view close to 160° [2]. Due to surgical limitations on implant size, current retinal prostheses only stimulate the central 15–20° field of view. Prototype systems range from 16 to 1550 electrodes [3–6], which is well below the resolution of the retina in this region, even if every electrode can create an independent pixel. If the entire camera image (between 40° and 60° field of view) is compressed to fit the central field, there will be a loss of resolution and miniaturization of objects, with a likely decrease in the quality of vision. Whereas, if only the central 15–20° field of view from the camera image is extracted and stimulated electrically, the visual information will be more organized and perceivable to the recipients. However, peripheral information will be lost, severely hampering mobility. Hence, there is a need for a specific image processing algorithm which could be used to overcome the loss of peripheral information due to the limited field of view.

Retinal prosthesis research can involve image processing in a number of different ways. Several studies have conducted simulated vision experiments with normal sighted volunteers performing reading or mobility tasks. The goal of these experiments was to test visual task performance as a function of pixel number, density and quality. These have been recently

 © 2010 IOP Publishing Ltd Printed in the UK

reviewed [7] and collectively, these studies suggest that 600–1000 electrodes are needed for functional vision. Another facet of image processing research related to retinal prostheses involves the conversion of the image data into a stimulus pattern that best conveys the desired visual perception. Asher *et al* [8] propose real-time image processing algorithms and transformations to simulate the different functions of the various cell layers and cell layer connections in the retina. They also propose a conversion method for transforming the visual information into electrical current patterns. Hallum *et al* [9] also propose a method for converting an image frame captured by the camera into low resolution modulated charge injections. Finally, image processing has been proposed to enhance certain features of an image that may be important to the user. Boyle *et al* [10] examined accentuating certain image features. One finding from this study suggested the utility of important maps, like the ones created by the bottom-up visual attention saliency model by Itti *et al* [11–15]. With this in mind, we propose an image processing algorithm based on the saliency detection model by Itti *et al* to find the important and salient regions in the entire image frame and cue subjects toward the direction of the salient region.

The retina along with higher visual processes guides visual attention in the visual cortex. Visual attention binds information from multiple parallel processes carrying motion, depth, color and form information in the visual cortex [16]. During visual search different information from these processes is combined first, and the output guides the attention deployment process [17, 18]. Computational models based on saliency detection have been used in computer vision and robotics to predict important areas in the visual field. A first model based on the feature integration theory was proposed by Koch and Ullman [19]. The model is a bottom-up saliency detection model that computes a saliency map by combining several basic features which undergo parallel processing. Based on this model, a bottom-up model based on primate vision was proposed by one of the authors of this paper [11–15]. This model (hereon referred to as the 'full model') forms the basis of many implementations of visual attention in robotics and artificial intelligence [20] and also forms the basis of the work proposed here. We propose an algorithm (hereon referred to as the 'new model') that is based on the full model, but with simplifications to increase efficiency, to allow execution on a portable processor. In this paper, we describe the new model in detail and verify that it can predict human gaze using a library of images and human observers.

## 2. Methods

### *2.1. New model algorithm*

The new model (figure 1) uses three information streams: color saturation, intensity and edge information. These information streams are extracted by converting the input image from the RGB color space to the HSI (hue–saturation–intensity) color space. This conversion can be done in various ways. The conversion for our algorithm was done using the function rgb2hsv in Matlab from Mathworks Inc. Nine scales of dyadic Gaussian pyramids [21] are created for the saturation (S), intensity (I) and edge (E) information by successively low pass filtering and down sampling by a factor of 2. Edge pyramids are created from the intensity stream based on Laplacian pyramid generation [22, 23]. For each level of the pyramid, the edge pyramid image is created as a point-by-point subtraction between the intensity image at that level and the interpolated intensity image from the next level.

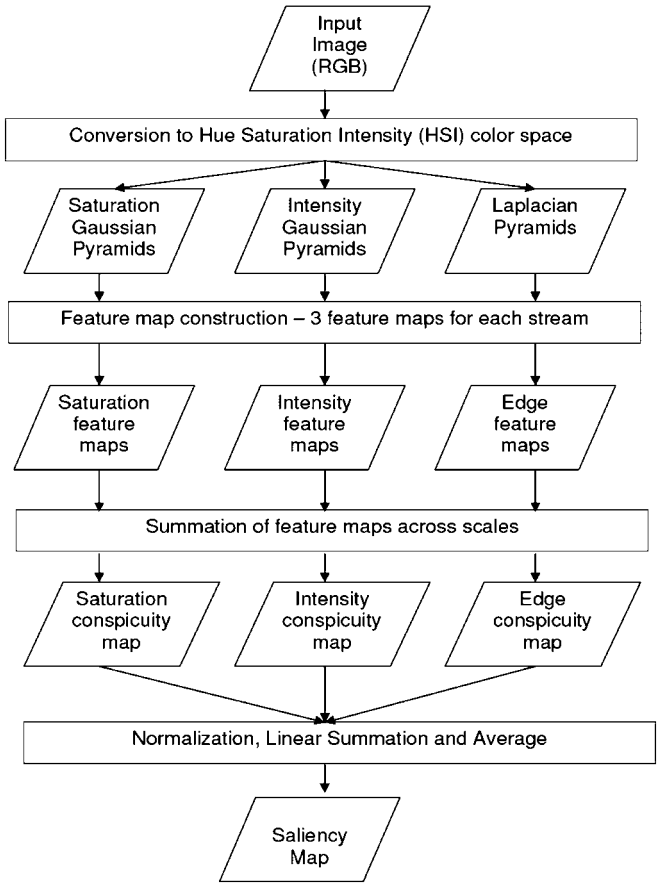



**Figure 1.** Diagram of the proposed algorithm.

Center–surround mechanisms observed in the visual receptive fields of the primate retina are then implemented computationally to create feature maps for each information stream. Center–surround interactions are modeled as the difference between the coarse and fine scales of the pyramids [11–14]. Feature maps are created from only four scales with the center scales '$c$' at levels (3, 4) and surround scales '$s$' at levels (6, 7) where the original image is at level 0 of the pyramid. For $c \in \{3, 4\}$ and $s = c + \delta$ where $\delta \in \{3, 4\}$ and $s < 8$, a set of three feature maps is created for each stream. Point-by-point subtraction between the values of the pyramids at the finer and coarser scales is carried out after interpolating the coarser scale to the finer scale using bilinear interpolation. Absolute values of the subtraction are calculated for the saturation and intensity streams and all feature-maps are decimated by dropping the appropriate number of pixels to be 1/16th the size of the original image:

$$S(c, s) = |S(c) - S(s)| \tag{1}$$

$$I(c, s) = |I(c) - I(s)| \tag{2}$$

$$E(c, s) = E(c) - E(s). \tag{3}$$

A linear summation across these feature maps for the different information streams forms the conspicuity maps—$S_c$ for color saturation, $I_c$ for intensity and $E_c$ for edge information:

$$S_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} S(c, s) \tag{4}$$

$$I_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} I(c, s) \tag{5}$$

$$E_c = \sum_{c=3}^{4} \sum_{s=c+3}^{c+4;s<8} E(c, s). \tag{6}$$

The conspicuity maps undergo normalization [11–15] referred to by the operator $\mathcal{N}$ in the equations. Normalization is an iterative process that promotes maps with a small number of peaks with strong activity and suppresses maps with many peaks of similar activity. Each conspicuity map is first normalized to a fixed range between 0 and 1. Thereafter, a two-dimensional difference of Gaussian filter (DoG) is convolved with the map iteratively. The output is summed with the original map and negative values are set to zero. The DoG filter results in the excitation of each pixel with inhibition from neighboring pixels. The DoG filter function is calculated as stated below:

$$\mathrm{DoG}(x, y) = \frac{0.5\,\mathrm{e}^{-(x^2+y^2)/(2\sigma_{\mathrm{ex}})}}{2\Pi\sigma_{\mathrm{ex}}^2} - \frac{1.5\,\mathrm{e}^{-(x^2+y^2)/(2\sigma_{\mathrm{inh}})}}{2\Pi\sigma_{\mathrm{inh}}^2}, \tag{7}$$

where $\sigma_{\mathrm{ex}} = 2\%$ and $\sigma_{\mathrm{inh}} = 25\%$ of the input image width.

Intensity and saturation conspicuity maps use three normalization iterations, and edge conspicuity maps use one normalization iteration. The number of iterations for normalization is chosen based on the computational load and pilot studies that examined different iterations of normalization and their effects on the maps. The three normalized conspicuity maps are linearly summed and their average forms the final salience map which again undergoes a three-iteration normalization. The region around the pixel with the highest grayscale value in the final salience map signifies the most salient region:

$$S = \mathcal{N}\left(\frac{\mathcal{N}(S_c) + \mathcal{N}(I_c) + \mathcal{N}(E_c)}{3}\right). \tag{8}$$

There are a few key differences between the two models which make the new model less computationally intensive compared to the full saliency model. The new model uses only 3 information streams for processing (versus 7 in the full model), 4 scales of Gaussian pyramids (versus 6), 18 feature maps (versus 42). Instead of using the two color opponent streams as found in the primate retina, the new model uses color saturation. Color saturation information will indicate purer hues with higher grayscale values and impure hues with lower grayscale values. One stream of edge information is used instead of four different orientation streams. For creating feature maps, the new model focuses on the coarser scales for center and surround which represent low spatial frequency information in the image.

### *2.2. DSP implementation*

The retinal prosthesis image processing system is designed to be a portable module worn on the body. For this reason, the module should be lightweight and compact with low power requirements so that it can operate for several hours on a small battery. Low power requirements may restrict the amount of computation that can be carried out by the image processing unit. With this in mind, for research purposes, the algorithm has been implemented on a Texas Instruments Imaging Developers Kit (IDK) TMS320 DM642 [24]. The DM642 chip is a 720 MHz fixed-point processor and the IDK is specifically designed to aid the development of image processing algorithms. The IDK is not a portable board (it includes many functions) but in general, DSPs are designed for low-power, portable applications, so the technological path is clear.

As a first step toward analyzing the computational speed of the new model, we implement the new model and full model (partial) on the DSP-IDK. The algorithms are modeled in Simulink from Mathworks Inc., and then ported to the DSP. For efficiency, filtering has been implemented using separable one-dimensional filters for both the models. Only the intensity stream of the full model is implemented (one of the seven streams of the full model). Fixed-point hardware can implement numbers in both fixed-point precision and floating-point precision. Both the algorithm implementations are in single-precision floating-point format and are not optimized.

### *2.3. Model validation using gaze data*

Gaze experiments were carried out with five human subjects after the approval of the Institutional Review Board at the University of Southern California. A signed informed consent was obtained from each participant of the study. Subjects were required to have English speaking and reading knowledge, be 18+ years of age, not have a history of vertigo, motion sickness or claustrophobia; cognitive or language/hearing impairments, and have a visual acuity of 20/40 or better with normal or corrected vision (with lenses). Visual acuity testing was carried out in the lab using a Snellen visual acuity eye chart.

Gaze data were acquired using an eye tracking system from Arrington Research, Inc., Scottsdale, AZ. This system consists of a Z800 3D Visor Head Mounted Display (HMD) with a diagonal field of view of 40°. Images on the HMD are displayed at a resolution of 800 × 600 pixels. The Viewpoint eye tracking software from Arrington Research recorded data at a frequency of 60 Hz using pupil tracking. Subjects were seated at a table with their head rested on a chin rest. A 12 point rectangular grid calibration process was used. Subjects were asked to look at the center of 12 different squares that would successively appear on the HMD screen. After this, as a measure of calibration, a test image consisting of a circle in the center of the screen was shown and subjects were asked to look at the center of the circle. Recording was not done until good calibration was obtained. Good calibration is defined as a rectangular grid mapped from the gaze points of the subjects when looking at the 12 squares. A set of 150 natural images

was displayed on the HMD with each image being shown for 3 s. The images consisted of outdoor and indoor environments. Subjects were instructed to freely gaze at the image. To avoid biasing the subjects, no other instructions were given. Between images, the test circle image was displayed for subjects to rest their eyes. However, after every three images, subjects were instructed to look at the center of the circle and the calibration at this point was noted. This helped to keep track of any calibration drift during the experiments. In post-processing, the recorded gaze point data were then corrected by any offset to get the new gaze point data, corrected for calibration drift.

The collected gaze data were filtered for fixations and saccades using custom fixation and saccade filtering software freely available on http://ilab.usc.edu as part of the Neuromorphic Vision Toolkit. Data were analyzed using gaze fixation points from the data set. Fixation data points may not account for drifts in eye movements. However, by taking a circular aperture around each fixation point during data analysis, effects of drifts as well as slight calibration offsets can be avoided.

Analysis of gaze data was carried out using methods used by Itti [25] and Peters *et al* [26] to analyze the contribution of bottom-up saliency to human eye movements. For each image, gaze data points from all subjects were pooled together for analysis. For the same set of 150 input images, salience maps created by the full model were used for a comparison of results with salience maps created by the new model.

*2.3.1. Analysis with the ratio of medians method [25].* $S_h$ is defined as the highest value of saliency within a circular aperture of diameter 5.6° centered at the fixation point. High values of $S_h$ indicate that the human observer fixated at a highly salient region.

$S_r$ is defined as the highest value of saliency within a 5.6° circular aperture, centered at a random point chosen from a uniform distribution.

$S_{max}$ is defined as the maximum value of saliency in the salience map of the image.

Each image will have approximately between 20 and 40 gaze points after combining gaze data from all subjects. The same number of points are randomly chosen from a uniform distribution to calculate $S_r$. To get a more accurate estimate for $S_r$, 100 sets of random points are used for each image, each generating an $S_r$ value. The median $S_{rm}$ of this set of $S_r$ values for each image is used for further analysis. Ratios $S_h/S_{max}$ and $S_{rm}/S_{max}$ and the medians for each of these are calculated. The ratio of these medians is then calculated. Higher ratios mean that saliency values around fixation points are greater than saliency values around random end points, showing that the model can predict human gaze locations in the image better than expected by chance.

*Image shuffling.* Shuffling is a control analysis where instead of using gaze points for the image in consideration, gaze points of another randomly chosen image are used. The ratio of medians analysis stated above is done using the saliency maps from one image and gaze data from the randomly chosen image. These results are then compared to the results when using the saliency maps and gaze data for the same image.

Differences between $S_h$ and $S_{rm}$ were evaluated using a statistical sign test with a significance level of 0.0001 for both the full and new models for the cases with and without shuffling. Also, the same statistical test was carried out between the $S_h$ values with and without shuffling to see if the $S_h$ values with shuffling are significantly less than the $S_h$ values without shuffling.

*2.3.2. Analysis using normalized scanpath salience (NSS) [26].* This method normalizes the salience map to have a zero mean and unit standard deviation. For each point corresponding to the fixation locations, the normalized salience value is extracted and the mean of all these extracted values is calculated. This mean is the normalized scanpath salience (NSS) value. If the NSS value is greater than zero, there is a greater correspondence between the salience maps and gaze fixation points than expected by chance. The NSS value of zero would mean there is no such correspondence and a value of less than zero would mean there is anti-correspondence between the salience maps and human fixations. To verify this in practice, chance values are calculated by creating a map with a uniform distribution at the same resolution as the saliency map instead of the actual salience map and calculating NSS in the same manner as stated above. We calculated the NSS values for all gaze data points by taking a region of diameter 5.6° around each fixation point in order to avoid any fixation drifts and minor calibration offset effects. Here again, random map generation was carried out 100 times for each image.

For both the full and new models, NSS values obtained using salience maps were compared to the NSS values obtained using random maps (paired *t*-test with a significance level of 0.0001).

## 3. Results

### 3.1. Saliency maps

Figure 2 shows the salience maps generated by the new and full models for the same input image. Figure 2(a) shows an example of an input image, the saturation, intensity and edge conspicuity maps and the final salience map created by the new model. Figure 2(b) shows the salience map computed by the full model for the same input image. On comparing the final saliency maps from the new and the full models, we can observe that the salient image areas (e.g. the curb, the plant, etc) are similar in the outputs of both models. The new model works with a lower resolution image (320 × 240 pixels) than the full model (640 × 480 pixels) resulting in coarser maps when compared to the full model.

Saturation and edge conspicuity maps (figure 2(a)) enhance objects with more saturated hues and darker colors and objects with prominent edges respectively whereas the intensity conspicuity map enhances objects with intensity contrast in the image frame. The process of creating the feature maps and normalizing them can lead to certain pixels not being enhanced in the final conspicuity map.

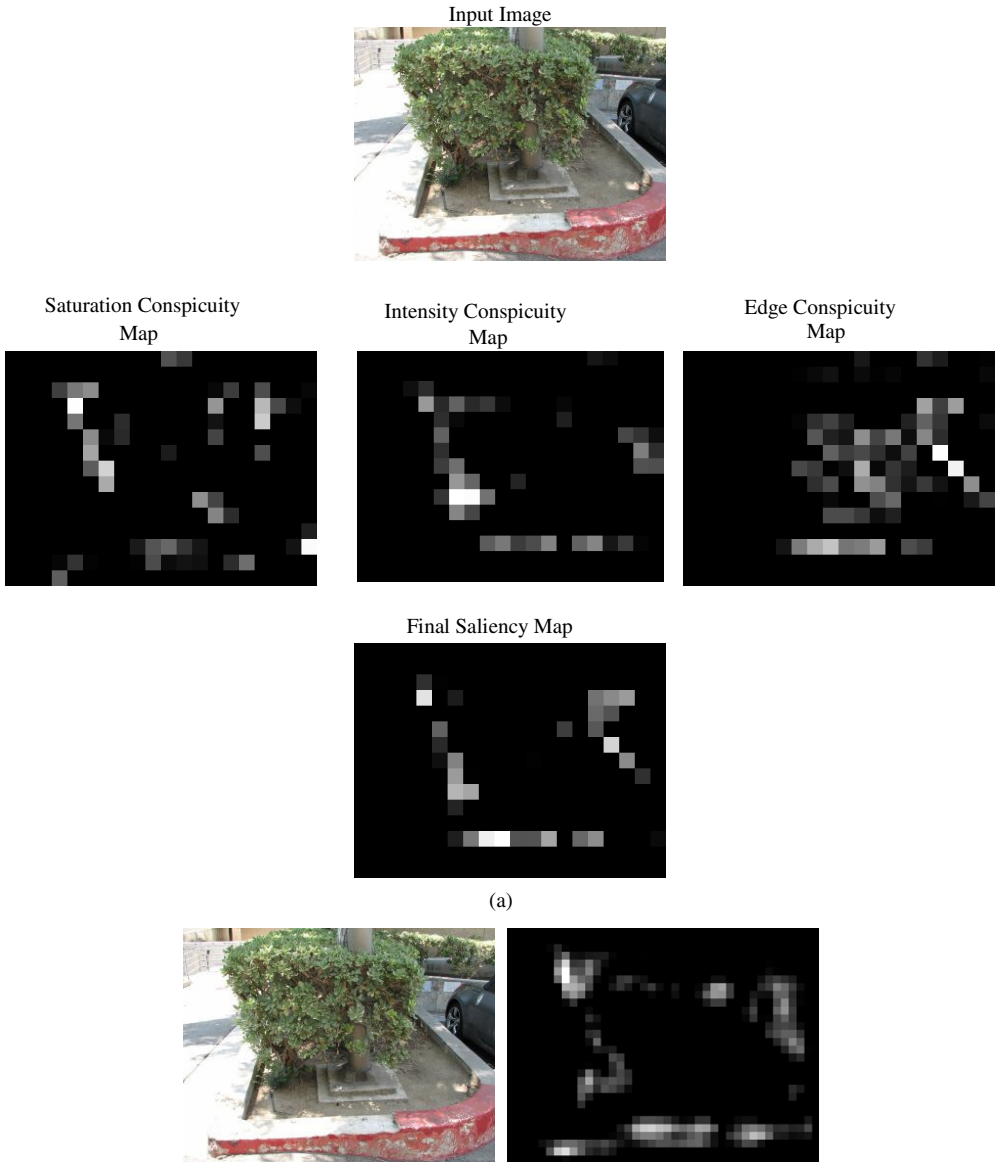



**Figure 2.** Salience maps created by the new model and the full model for the same input image: (a) conspicuity maps for saturation, intensity and edge along with the final salience map created by the new model for an example input image; (b) final salience map created by the full model for an example input image.

### *3.2. DSP implementation results*

The various modules of each model and the time required to process one frame are stated in table 1. As stated earlier, for the full model, only the intensity stream which is one of seven different streams has been implemented on the DSP.

Execution time results from table 1 show that a single image frame takes 0.84 s to be processed by the new model whereas 14% of the old model (only the intensity stream) takes 1.53 s to process the same image. This shows that the implementation of the new model is computationally more efficient. The estimated time for the full model to execute one frame can be calculated by multiplying the time for the intensity stream execution by a factor of 7. This is because the intensity stream is one of seven similar streams in terms of the computational complexity in the original saliency algorithm. This implies that the implementation of the new model is approximately ten times faster than the implementation of the full model.

Optimization can lead to better results as can improvements in processor speed and power consumption. However, in wearable computing systems, increased algorithm efficiency will always translate into lower power consumption. Even the unoptimized implementation of the new model can execute in less than 1 s, which is a reasonable response time to a user request for information.

### *3.3. Model validation using gaze data*

*3.3.1. Analysis with the ratio of medians method [25].* Figure 3 shows two examples of input images with salience maps from the new model and the full saliency model. Points in images

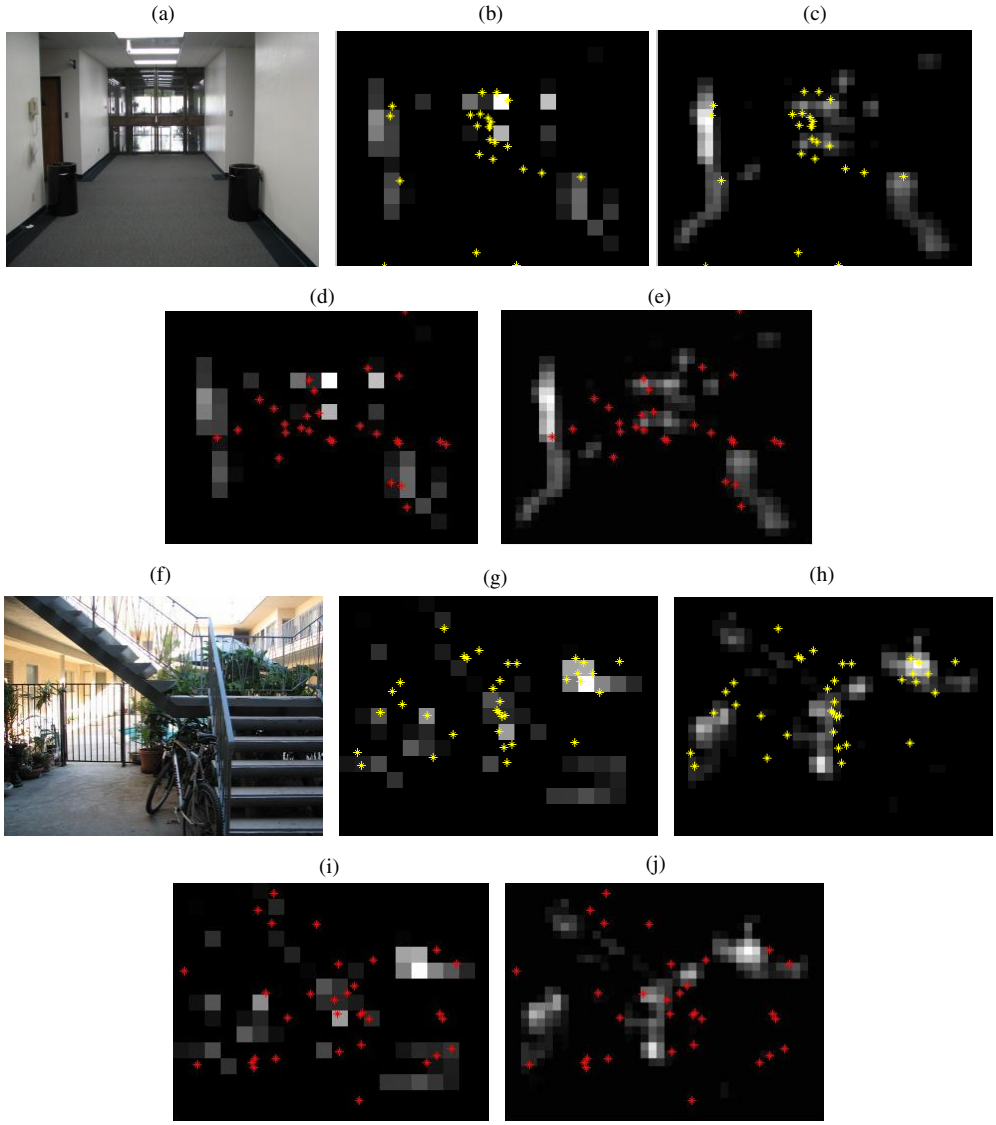



**Figure 3.** Input image (a and f), salience maps from new model (b and g) and full saliency model (c and h) with the dots depicting gaze fixation points, and salience maps from the new model (d and i) and full saliency model (e and j) with dots depicting data points after shuffling.

**Table 1.** Time in seconds for computation of different modules in the new model and intensity stream of the full model on the TMS320 DM642 DSP.

| Functions | New model implementation (time in seconds) | Intensity stream from the full model implementation (time in seconds) |
|---|---|---|
| YCbCr → RGB | 0.1250 | – |
| RGB → HSI | 0.0320 | – |
| Gaussian pyramids (intensity and saturation for new model) | 0.0647 | 0.0695 |
| Laplacian pyramids | 0.0303 | – |
| Center–surround maps | 0.0027 | 0.0158 |
| Normalization function | – | 0.6062 |
| (at different scales) | – | 0.0757 |
| | 0.0096 | 0.0113 |
| Entire algorithm (s) | 0.8416 | 1.5373 |
| Frames/second | 1.1882 | 0.6505 |

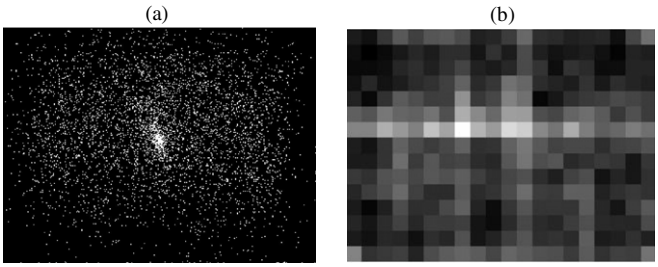



**Figure 4.** Gaze distribution of all subjects over all images (a) and the average salience map from the salience maps of all images (b) in the entire data set.

**Table 2.** Data analysis of human gaze data with salience maps created by the new model and the full saliency model using the ratio of medians method.

| | Median ($S_h/S_{max}$) | Median ($S_{rm}/S_{max}$) | Ratio of medians | Sign test ($S_h$ and $S_{rm}$) |
|---|---|---|---|---|
| New model | 0.3647 | 0.1020 | 3.5769 | $p < 0.0001$ |
| Full model | 0.4352 | 0.2457 | 1.7714 | $p < 0.0001$ |
| Image shuffling | | | | |
| New model | 0.2275 | 0.1059 | 2.1481 | $p < 0.0001$ |
| Full model | 0.3256 | 0.2511 | 1.2970 | $p < 0.0001$ |

(b), (c), (g) and (h) depict gaze fixation points from human subject data and the points in images (d), (e), (i) and (j) depict data points obtained by shuffling (gaze points from another image).

Table 2 shows analysis of the full and the new models for the actual gaze data and randomly distributed gaze points. Both the full and the new models have ratios which are significantly above chance (sign test with a significance level of 0.0001 carried out between the $S_h$ and $S_{rm}$, chance = 1) indicating that both models predict better than chance where human observers will look. The ratio of medians calculated by the new model is higher than the full model which shows that the new model outperforms the full model in this case. The maps from the full model are slightly denser than the maps from the new model as seen in the comparison between figures 3(i) and (j). This results in the overall median values of the full model being greater than the new model.

The shuffled image analysis results are also shown in table 2. A statistical sign test with a significance level of 0.0001 between the $S_h$ values with and without shuffling indicates that the $S_h$ values without shuffling are significantly higher than the $S_h$ values with shuffling. The shuffled analysis shows that the median values and ratios are lower than when the saliency maps and gaze data correspond, but the ratios are statistically greater than one, that is, better than chance. This discrepancy can be explained by the center-bias effect present in the average salience map of all images as well as in the gaze data of subjects. When looking at unfamiliar images, subjects often start looking at the center and then proceed to examine the peripheral areas. Subjects are asked to look at the center of a test image after every three images for calibration purposes as mentioned before which could also add to their initial fixation being centrally biased. Finally, due to potential photographer bias (having interesting objects in the center of the image), the average of all the salience maps in the input image data set also has a center bias. Figure 4 shows the center-bias in the gaze data as well as the average salience map.

**Table 3.** Data analysis of human gaze data with salience maps created by the new model and the full saliency model using the ratio of medians method for the image data set after removing images with a center bias in the gaze data and/or the salience maps.

| | Median ($S_h/S_{max}$) | Median ($S_{rm}/S_{max}$) | Ratio of medians | Sign test ($S_h$ and $S_{rm}$) |
|---|---|---|---|---|
| New model | 0.3490 | 0.1020 | 3.4231 | $p < 0.0001$ |
| Full model | 0.4278 | 0.2519 | 1.6985 | $p < 0.0001$ |

To investigate the finding that saliency and gaze were correlated even with image shuffling, a center-bias analysis of the gaze points from subjects as well as the average salience map was done. Based on Tatler's analysis [27], for each image, the number of gaze points falling into the central 15° was counted and compared to the number of gaze points falling into the rest of the image areas which are referred to as peripheral areas. If the number of gaze points in the central region was greater than the number in the peripheral regions, there was a center bias in gaze data. Calculations show that 26% of the images used in our study have a center bias in the subject gaze data. Similarly, the number of pixels whose grayscale level is the maximum value of the average salience map is calculated in the central and peripheral areas of the average salience map. If the number of such maximum grayscale valued pixels is greater in the center than in the periphery, there is said to be a center bias. Figure 4(b) shows the average saliency data from all images, indicating central bias. The bias in the subject gaze while viewing these unseen natural images and the bias in the salience maps due to the photographer bias may be a reason behind the ratio of medians being greater than 1 even with image shuffling. In general, if the combination of the shuffled gaze data set and the salience map is such that both have an overlap, the ratio for such combinations will be greater than 1.

Analysis was repeated after removing images with a central bias. Table 3 shows these results. The median values for $S_h/S_{max}$ and $S_{rm}/S_{max}$ are very close to those obtained with the entire set of images including the ones with a center bias. As before, a sign test with a significance level of 0.0001 carried out between the $S_h$ and $S_{rm}$ shows that $S_h$ values are significantly higher than $S_{rm}$ values.

*3.3.2. Analysis using normalized scanpath salience (NSS) [26].* Figure 5 shows two examples of an input image with salience maps from the new model and the full saliency model. The figure also shows a random map created from a uniform distribution at the same resolution as the salience maps for the new and full models. The dots in the salience and random maps represent the gaze fixation points of human observers.

The NSS results are shown in table 4 for the new model as well as the full model. NSS values for both models with salience maps are greater than zero whereas the NSS value for the random model is close to zero. The results for the analysis on the data set of images after removing images with a gaze or salience map center bias are shown in table 5.

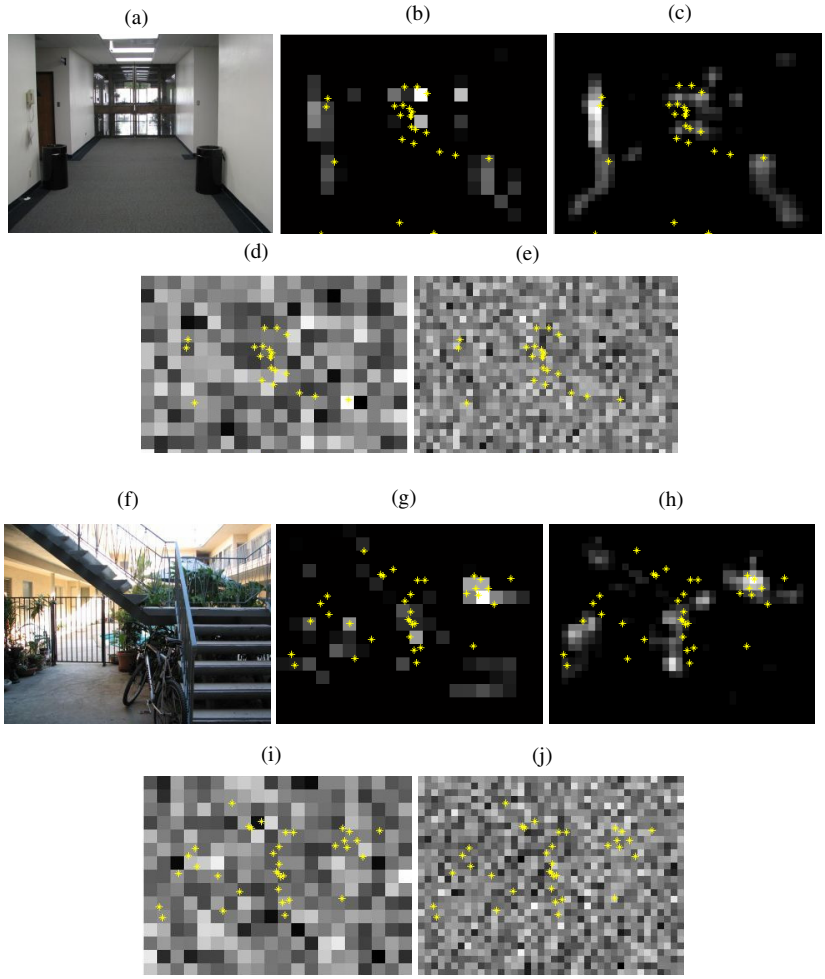



**Figure 5.** Input image (a and f), salience maps from the new model (b and g) and the full saliency model (c and h); uniform distribution random map with the dots depicting the gaze fixation points of the human subjects for the new model (d and i) and for the full model (e and j).

**Table 4.** Data analysis of human gaze data with salience maps created by the new model and the full model using the normalized scanpath salience method.

| | For salience map NSS ± SEM | For random map NSS ± SEM | Paired *t*-test |
|---|---|---|---|
| New model | $0.4310 \pm 0.0113$ | $-4.813 \times 10^{(-4)} \pm 0.0005$ | $p < 0.0001$ |
| Full model | $0.4758 \pm 0.0098$ | $-5.077 \times 10^{(-4)} \pm 0.0005$ | $p < 0.0001$ |

**Table 5.** Data analysis of human gaze data with salience maps created by the new model and the full model using the normalized scanpath salience method for the image data set after removing images with a center bias in the gaze data and/or the salience maps.

| | For salience map NSS ± SEM | For random map NSS ± SEM | Paired *t*-test |
|---|---|---|---|
| New model | $0.4153 \pm 0.0104$ | $1.2311 \times 10^{(-4)} \pm 0.0006$ | $p < 0.0001$ |
| Full model | $0.4746 \pm 0.0093$ | $1.9836 \times 10^{(-4)} \pm 0.0005$ | $p < 0.0001$ |

For both cases, a paired *t*-test with a significance level of 0.0001 shows that the NSS values obtained using salience maps are significantly different than the NSS values obtained using random maps, meaning there is greater correspondence between salient regions detected by the salience maps and human fixations than expected by chance.

## 4. Discussion

We present a computationally efficient model of bottom-up saliency detection based upon an earlier saliency model [11–15]. Good correspondence is noted when comparing regions that the algorithm predicts as salient to regions gazed

at by human subjects when looking at a set of images. Also comparing the salient regions to random gaze points shows that the model predicts salient regions at a rate better than what would be expected by chance. We have validated our algorithm with sighted observers viewing images on a computer screen while seated. The subjects are shown unfamiliar scenes to limit unbiased gaze patterns. Using a set of 150 images and gaze data from five subjects, results show comparable performance between the new and the full models. An unoptimized implementation on the TMS320 DM642 DSP shows that the proposed model can process at a rate of 1 frame per second which is approximately ten times faster than the full model. Since this algorithm is eventually hoped to be run on a wearable computing platform, efficiency is a critical factor.

The model is proposed as the core of an image processing algorithm designed to provide visual prosthesis patients with information about the areas outside the visual field of the implant. Such an algorithm could be utilized in a number of ways. During navigation and ambulation, the user might want to know about obstacles or signs (for example, an exit sign). Other times, the user might be searching for an object of interest. While it is possible to design specific algorithms tailored to each task, there are advantages to a bottom-up approach. Unlike top-down algorithms that require *a priori* information, bottom-up algorithms do not require any training. Also, a bottom-up algorithm may allow the user to identify objects and understand surroundings using remnant vision and contextual cues. Nevertheless, it is possible that a top-down algorithm may be needed for specific tasks such as objective recognition, particularly where vision is very poor. Frintrop *et al* proposed a saliency implementation based on ten information streams and a five level image pyramid scheme for robotics [20]. Walther *et al* proposed a bottom-up implementation based on the full model with an added feedback module to detect the extent of an attended object [28]. Both the groups combined their bottom-up implementations by using top-down information based on feature detection and/or object recognition with the salient regions [29, 30].

To be effective for a retinal prosthesis implant patient, more work is required to understand what functions are important for these patients. Training will also be required to best utilize information provided by the algorithm. Also, it is unclear if patients can learn to take advantage of the additional information or if they will prefer to receive unfiltered video data and make their own judgments about object importance. Task-dependent processing may be the best approach. For obstacle avoidance and route planning, visually impaired people are likely to be more interested in large objects obstructing their path versus small details of the environment around them. In such a case, a saliency algorithm may be adequate. For an object detection task, smaller details may be important discriminating clues to aid successful task completion and a top-down object recognition algorithm may be required. Any additional algorithm will require more computing power so additional benefit will come at a cost. The extra computing load can be limited by applying object recognition only in the small region identified as salient. This would also eliminate the need for the entire image frame to be processed in smaller parts by the object recognition algorithm to find the various objects.

In summary, a computationally efficient image processing algorithm has been analyzed and identifies parts of an image that human observers also deem salient. This algorithm has the potential to enhance low vision, particularly when visual field is restricted. When used with a retinal prosthesis, the algorithm can be implemented on the retinal prosthesis' existing camera and wearable computing platform. Critical questions remaining to be answered by human testing include how quickly people learn to utilize the algorithm and the benefit provided.

## Acknowledgments

This research was performed at the Biomimetic Microelectronics Systems Engineering Research Center. This material is based on the work supported by the National Science Foundation under grant no EEC-0310723. This work was also supported (LI) by grants from the National Science Foundation, the Defense Advanced Projects Agency and the Army Research Office. The views, opinions and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Agency or the Department of Defense. We would also like to thank Dr Armand Tanguay and his student Benjamin McIntosh for the Arrington Research equipment to record gaze data and David Berg at iLab for his help with the saccade and fixation filtering software.

## References

[1] Margalit E and Sadda S R 2003 Retinal and optic nerve diseases *Artif. Organs* **27** 963–74
[2] Kolb H, Fernandez E and Nelson R 2002 *Webvision: The Organization of the Retina and the Visual System*
[3] de Balthasar C *et al* 2008 Factors affecting perceptual thresholds in epiretinal prostheses *Invest. Ophthalmol. Vis. Sci.* **49** 2303–14
[4] Yanai D *et al* 2007 Visual performance using a retinal prosthesis in three subjects with retinitis pigmentosa *Am. J. Ophthalmol.* **143** 820–7
[5] Gekeler F, Sachs H, Szurman P, Guelicher D, Wilke R, Reinert S, Zrenner E, Bartz-Schmidt K and Besch D 2008 Surgical procedure for subretinal implants with external connections: the extra-ocular surgery in eight patients *ARVO 2008 Annual Meeting* ARVO Abstract-4049
[6] Humayun M S 2009 Preliminary results from Argus II Feasibility study: s 60 electrode epiretinal prosthesis *ARVO 2009 Annual Meeting* ARVO Abstract 4744
[7] Dagnelie G 2008 Psychophysical evaluation for visual prosthesis *Ann. Rev. Biomed. Eng.* **10** 339–68
[8] Asher A *et al* 2007 Image processing for a high-resolution optoelectronic retinal prosthesis *IEEE Trans. Biomed. Eng.* **54** 993–1004
[9] Hallum L E, Cloherty S L and Lovell N H 2008 Image analysis for microelectronic retinal prosthesis *IEEE Trans. Biomed. Eng.* **55** 344–6

[10] Boyle J R, Maeder A J and Boles W W 2008 Region-of-interest processing for electronic visual prosthesis *J. Electron. Imaging* **17** 013002
[11] Itti L, Koch C and Niebur E 1998 A model of saliency-based visual attention for rapid scene analysis *IEEE Trans. Patt. Anal. Mach. Intell.* **20** 6
[12] Itti L 2000 Models of bottom-up and top-down visual attention *PhD Thesis* California Institute of Technology, Pasadena p 216
[13] Itti L and Koch C 2000 A saliency-based search mechanism for overt and covert shifts of visual attention *Vis. Res.* **40** 1489–506
[14] Itti L and Koch C 2001 Computational modelling of visual attention *Nat. Rev. Neurosci.* **2** 194–203
[15] Itti L and Koch C 2001 Feature combination strategies for saliency-based visual attention systems *J. Electron. Imaging* **10** 161–9
[16] Kandel R, Schwartz J and Jessel T 2000 *Principles of Neural Science* 4th edn (New York: McGraw-Hill)
[17] Wolfe J M, Cave K R and Franzel S L 1989 Guided search: an alternative to the feature integration model for visual search *J. Exp. Psychol. Hum. Percept. Perform.* **15** 419–33
[18] Wolfe J M 1994 Guided search 2.0: a revised model of visual search *Psychon. Bull. Rev.* **1** 202–38
[19] Koch C and Ullman S 1985 Shifts in selective visual attention: towards the underlying neural circuitry *Hum. Neurobiol.* **4** 219–27
[20] Frintrop S 2005 *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search* (Germany: University of Bonn)
[21] Greenspan H, Belongie S, Goodman R, Perona P, Rakshit S and Anderson C H 1994 Overcomplete steerable pyramid filters and rotation invariance *IEEE Computer Vision and Pattern Recognition (Seattle, Washington)*
[22] Burt P J and Adelson E H 1983 The Laplacian pyramid as a compact image code *IEEE Trans. Commun.* **31** 532–40
[23] Adelson E H, Anderson C H, Bergen J R, Burt P J and Ogden J M 1984 Pyramid methods in image processing *RCA Engineer* **29** pp 33–41
[24] Texas Instruments, *TMS320DM642 Video/Imaging Fixed-Point Digital Signal Processor (Rev. L)*
[25] Itti L 2005 Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes *Vis. Cogn.* **12** 1093–123
[26] Peters R J, Iyer A, Itti L and Koch C 2005 Components of bottom-up gaze allocation in natural images *Vis. Res* **45** 2397–416
[27] Tatler B W 2007 The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions *J. Vis.* **7** (4) 1–17
[28] Walther D and Koch C 2006 Modeling attention to salient proto-objects *Neural Netw.* **19** 1395–407
[29] Frintrop S, Backer G and Rome E 2005 Goal-directed search with a top-down modulated computational attention system *Proc. Ann. Meeting of the German Association for Pattern Recognition (DAGM'05) (Wien, Austria)*
[30] Walther D 2006 Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics *PhD Thesis* California Institute of Technology

# Beobot 2.0: Cluster Architecture for Mobile Robotics

**Christian Siagian, Chin-Kai Chang, Randolph Voorhies, and Laurent Itti**
*Department of Computer Science, University of Southern California, Los Angeles, California 90089*
*e-mail: siagian@usc.edu, chinkaic@usc.edu, voorhies@usc.edu, itti@pollux.usc.edu*

Received 6 April 2010; accepted 23 October 2010

With the recent proliferation of robust but computationally demanding robotic algorithms, there is now a need for a mobile robot platform equipped with powerful computing facilities. In this paper, we present the design and implementation of Beobot 2.0, an affordable research-level mobile robot equipped with a cluster of 16 2.2-GHz processing cores. Beobot 2.0 uses compact Computer on Module (COM) processors with modest power requirements, thus accommodating various robot design constraints while still satisfying the requirement for computationally intensive algorithms. We discuss issues involved in utilizing multiple COM Express modules on a mobile platform, such as interprocessor communication, power consumption, cooling, and protection from shocks, vibrations, and other environmental hazards such as dust and moisture. We have applied Beobot 2.0 to the following computationally demanding tasks: laser-based robot navigation, scale-invariant feature transform (SIFT) object recognition, finding objects in a cluttered scene using visual saliency, and vision-based localization, wherein the robot has to identify landmarks from a large database of images in a timely manner. For the last task, we tested the localization system in three large-scale outdoor environments, which provide 3,583, 6,006, and 8,823 test frames, respectively. The localization errors for the three environments were 1.26, 2.38, and 4.08 m, respectively. The per-frame processing times were 421.45, 794.31, and 884.74 ms respectively, representing speedup factors of 2.80, 3.00, and 3.58 when compared to a single dual-core computer performing localization. © 2010 Wiley Periodicals, Inc.

## 1. INTRODUCTION

In the past decade, researchers in the field of mobile robotics have increasingly embraced probabilistic approaches to solving hard problems such as localization (Fox, Burgard, Dellaert, & Thrun, 1999; Thrun, Fox, & Burgard, 1998; Thrun, Fox, Burgard, & Dellaert, 2000), vision (Heitz, Gould, Saxena, & Koller, 2008; Wu & Nevatia, 2007), and multirobot cooperation (Fox, Burgard, Kruppa, & Thrun, 2000; Thrun & Liu, 2003). These algorithms are far more sophisticated and robust than the previous generation of techniques (Brooks, 1986; Maes & Brooks, 1990; Pomerleau, 1993). This is because these contemporary techniques can simultaneously consider many hypotheses in forms of multimodal distributions. Because of that, however, they are also far more computationally demanding. For example, a visual recognition task, in which we need to compare the current input image captured by a camera against a large database of sample images (Bay, Tuytelaars, & Gool, 2006; Lowe, 2004; Mikolajczyk & Schmid, 2005), requires not only that robust visual features be extracted from the input image—which already is a computationally demanding task—but also that these features be matched against those stored in the database—an even more demanding task when the database is large. As a point of reference, comparing two $320 \times 240$ images using scale-invariant feature transform (SIFT) features (Lowe, 2004) can take 1–2 s on a typical 3-GHz single-core machine. To be able to run such algorithms in near real time, we need a mobile robot equipped with a computing platform significantly more powerful than a standard laptop or desktop computer.

However, existing indoor and/or outdoor mobile robot platforms commercially available to the general research community still appear to put little emphasis on computational power. In fact, many robots, such as the Segway RMP series (Segway, Inc., 2009), have to be separately furnished with a computer. On the other hand, robots that come equipped with multiple onboard computers either do not use the most powerful computers available today [e.g., the Seekur (MobileRobots, Inc., 2009), which relies on the less powerful PC/104 standard] or have fewer computers (e.g., Carnegie Mellon University Robotics Institute, 2009; Willow Garage, 2009) than our proposed solution.

Before describing the design and implementation of our robot, in Section 1.1 we survey the current trends in the mobile robot market and identify the most desirable features in an ideal research robot (aside from our central requirement of powerful computational facilities). Note that some of the robots discussed below may no longer be available (or may never have been) to the general public. We include them nonetheless for completeness of our analysis.

We then describe our main contribution in Section 1.2, the design and implementation of our proposed platform, Beobot 2.0, a powerful mobile robot platform equipped with a cluster of 16 2.2-GHz processing cores. Our robot uses compact Computer on Module (COM) processors with modest power requirements, thus accommodating

various robot design constraints while still satisfying the requirement for computationally intensive algorithms.

Our complete design specifications, including supplier and cost information for almost all the materials, are freely available on the Internet (Siagian, Chang, Voorhies, & Itti, 2009). As the manufacturing, assembly, and machining details are available online, in this paper we focus on (1) the design decisions we made, the implementational issues we faced, and how we resolved them, and (2) experimental testing of the robot in diverse tasks.

### 1.1. Current Mobile Robot Platforms

In the current state of robotics, researchers utilize a variety of mobile robots, from the commercially available (iRobot Corporation, 2009b; MobileRobots, Inc., 2009; Willow Garage, 2009) to the custom made (Carnegie Mellon University Robotics Institute, 2009). These robots are built for many different environments, such as underwater (iRobot Corporation, 2010; USC Robotics, 2009), aerial (Finio, Eum, Oland, & Wood, 2009; He, Prentice, & Roy, 2008), and land (Quigley & Ng, 2007; Salichs, Barber, Khamis, Malfaz, Gorostiza, et al., 2006). Here we focus on land robots of a size close to that of an adult human that can traverse most urban environments, both indoors and outdoors, and for considerable distances. In addition, it is versatile enough for research in many subfields such as localization/navigation, human–robot interaction, and multirobot cooperation.

Furthermore, we primarily focus on sites that are similar to a college campus setting, which is mostly paved with some rough/uneven roads, not terrains that one would see in combat zones. Nowadays, because there is a concerted effort by most governments to make pertinent locations accessible to the disabled (using wheelchairs), legged robots [from the small QRIO and AIBO by Sony (Sony Entertainment Robot Europe, 2009) to the human-sized Honda Asimo (American Honda Motor Co., Inc., 2009)] are no longer a must. A wheeled platform would suffice for the target environments. However, the ability to traverse reasonably sloped terrain (about 10 deg) should also be expected. Also, some form of weather protection in the outdoors is essential. Although the robot is not expected to operate in all kinds of harsh weather (pouring rain, for example), like Seekur and Seekur jr. by MobileRobots (MobileRobots, Inc., 2009) and IRobot's PackBot (iRobot Corporation, 2009a), it should nevertheless be able to handle most reasonable conditions.

An overall size that is close to that of an adult human is ideal because the robot would be small enough to go through narrow building corridors and yet large enough to travel between buildings in a timely manner. And thus we exclude small robots such as the Khepara (AAI Canada, Inc., 2009) or large robotized cars such as the entries to the DARPA Grand Challenge. Smaller robots such as the Roomba (iRobot Corporation, 2009b) and Rovio (Evolution Robotics, Inc., 2009) and slightly larger ones such as the Pioneer (MobileRobots Inc., 2009) are also excluded because of their lower payload capacity, which limits the amount of computing that can be carried to a single laptop.

Aside from mobility, a few other important features contribute to the usability of the robot. They are battery life, sensors, interfaces, and available software. An ideal battery system would be one that enables the user to run for a whole day without having to recharge. The two factors that matter here are the total charge carried and the amount of charge required to operate the robot. The latter is dictated by the total weight of the robot and power consumption of the computers and devices. These requirements should be decided first. On that basis, the former can then be adjusted by selecting the proper battery system (type and quantity).

There are different types of available batteries: NiCd, NiMH, sealed lead acid (SLA), and lithium based. The trade-off is that of cost, dimensions, and durability. For one, SLA batteries are the most economical (in terms of cost-to-charge ratio), widely available when it comes time to replace them, and robust as they are easy to maintain and long lasting. However, SLA batteries have low charge-to-weight as well as charge-to-volume ratios compared to, particularly, the lithium-based technologies. Lithium batteries are lighter and more compact for a comparable amount of charge (National Institute of Standards and Technology, 2009). However, these types of batteries are much more expensive and fragile than the very rugged SLAs. If lithium batteries are not handled carefully, for example by not using protective circuits, they can explode. Although for the size of robot we are considering battery weight is not as much an issue, note that volume would still matter in terms of placement.

It is important to have easy access to the battery compartment so that we do not have to unscrew or disassemble components in order to charge the batteries. If the batteries can be charged rapidly, within an hour or two, an even better option would be to be able to do so without having to remove them from the robot, instead using a docking station or a wall plug-in outlet. If rapid recharge is not available, a feature to hot swap the batteries as in Willow Garage (2009) to avoid shutting down computers in the switching process would be convenient.

For a platform to be applicable for a wide range of robotics and vision research, most commercial robots are furnished with a variety of sensors and manipulation tools such as a robot arm and also provide avenues for future expansion. When selecting a sensor, we look for compact, light, low-power devices that exhibit high accuracy and high update rates. Popular sensors such as laser range finders, sonar rings, cameras, inertial measurement units (IMU), global positioning systems (GPS), and compasses should be considered as potential accessories. For a camera in particular, negligible latency is a must. After a number of experiments, we found that Firewire (IEEE-1394) cameras were the best in minimizing delays, more so than Internet

protocol (IP) or universal serial bus (USB) cameras. In addition, related features such as pan–tilt–zoom, autofocus, image stabilization, low-light capability, and wide-angle or omnidirectional viewing setup are also made available by various companies.

As for sensor expansion, aside from anticipating the future extra payload, it is also important to have many accessible USB ports placed throughout the body of the robot and USB-to-serial converters for serial devices, as well as several microcontrollers that can preprocess slower input signals.

Another important feature is having multiple types of user interface. For example, USB inputs are useful to connect a keyboard and monitor to the computers in the robot to allow for hardware and operating system reconfiguration. In addition, many robots have reasonably sized (15–25 cm) full red–green–blue (RGB) liquid crystal display (LCD) monitors for visualization of the robot's state during test runs. Furthermore, wireless network connections for remote secure shell (SSH) logins give us additional flexibility to allow for safe and faster algorithm on-site debugging. At the same time, the robot can use the external connection to access outside network or Internet resources, which can be useful in some scenarios. A related feature in this category is a standard radio frequency (RF) remote controller for stopping the robot whenever autonomous driving starts to fail. Furthermore, most robots (Carnegie Mellon University Robotics Institute, 2009; MobileRobots, Inc., 2009) are equipped with large kill switches to stop the flow of power to the motors.

In addition to the hardware-related aspects, robotic companies also provide software libraries to conveniently access all the included devices and monitor low-level states such as battery charge and temperature. Some companies (MobileRobots, Inc., 2009) provide further value additions such as mapping and navigation tools and even a full-blown simulation environment. We list the common software offerings in Section 4, where we describe our freely available toolkit (Itti, 2009).

### 1.2. Our Approach

Beobot 2.0 is the next iteration of the Beobot system developed in our lab (Chung, Hirata, Mundhenk, Ng, Peters, et al., 2002). The original Beobot integrated two full-sized, dual-CPU motherboards for a total of four 1-GHz processors. For Beobot 2.0, we use eight dual-core COM systems. Each COM measures just 125 × 9.5 × 18 mm and nominally consumes only 24 W of power. Nonetheless, with a 2.2-GHz dual-core processor, a COM has the computing power equivalent to current dual-core laptop systems. Despite this state-of-the-art computing platform, we have managed to keep the overall cost of our research-level, cluster-based mobile robot to under $25,000 (detailed in Siagian et al., 2009).

One aspect of a COM system to underscore here is the ease with which its components can be upgraded. Because the input and output signals are routed through just two high-density connectors, one need only remove the current module and replace it with an upgraded one. Thus, as more and more powerful processors become available, Beobot 2.0's computer systems can keep pace, making it somewhat more resistant to the rapid obsolescence that is characteristic of computer systems. The ability to keep pace with processor technology is important because robotic algorithms are expected to continue to evolve and become ever more complex, thus requiring commensurate levels of computing power.

Beobot 2.0's computer system is mounted on an electric wheelchair base (Figure 1), with an overall size that is close to that of a human. This allows the robot to navigate through corridors and sidewalks and creates an embodiment that is ideal for interacting with people. We assume that the majority of these pertinent locations are wheelchair accessible, as required by law. We believe that even with this locomotion limitation, there are still enough physically reachable locations to perform comprehensive real-world experiments. Figure 1 shows the finished robot.

The rest of the paper is organized as follows: first, we describe the electrical system in Section 2 and then the mechanical system in Section 3. Section 4 goes into the details of our software library, highlighting the advantage of implementing a computing cluster in robotics research.

In Section 5 we examine the robot on various important operational aspects, the most important of which is computational speed/throughput, to demonstrate how one could benefit from such a complex computing cluster architecture. We test Beobot 2.0 using three benchmark algorithms. One is the popular SIFT (Lowe, 2004) object recognition. The second is a distributed saliency algorithm (Itti, Koch, & Niebur, 1998), which models the visual attention system of primates. The algorithm operates on a very large image of 4,000 × 4,000 pixels and returns the most salient parts of the image. The last one is a vision localization system by Siagian and Itti (2009) that requires the system to compare a detected salient landmark input with a large landmark database obtained from previous visits. All of these algorithms are part of the Vision Toolkit, available freely online (Itti, 2009), which also houses Beobot 2.0's software control architecture, including obstacle avoidance (Minguez & Montano, 2004) and lane following (Ackerman & Itti, 2005). We then summarize our findings (in Section 6) and what we have learned through the process of building this robot.

## 2. ELECTRICAL SYSTEM DESIGN AND IMPLEMENTATION

Figure 2 presents an overview of the electrical system. On the right-hand side of the figure, there are two baseboards, each housing four COM Express modules (explained in

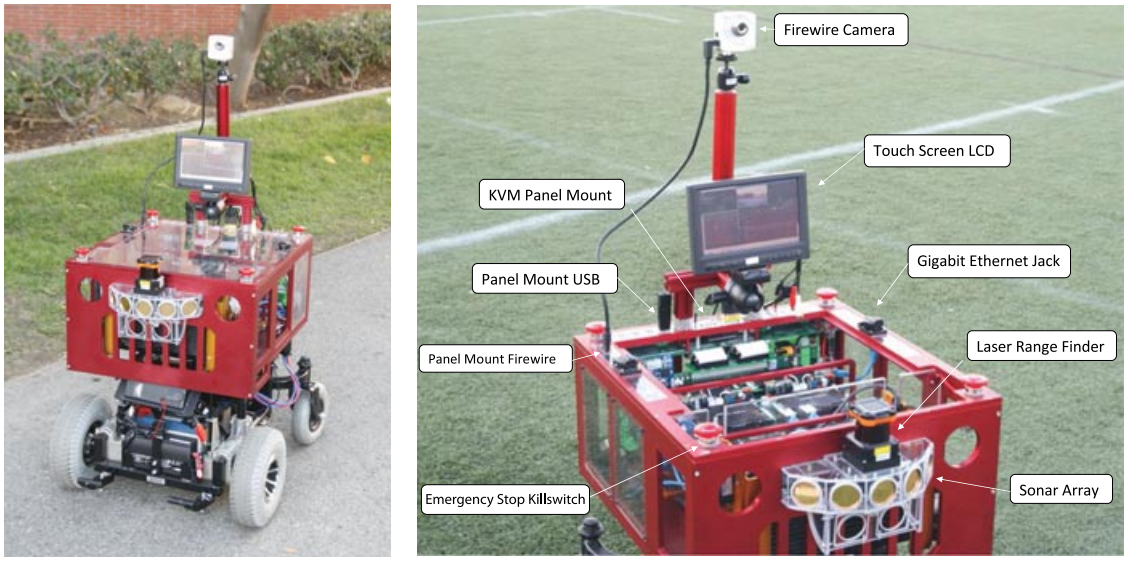



**Figure 1.** Various features of Beobot 2.0. Beobot 2.0 utilizes an electric wheelchair platform to carry a high-performance computing cluster of 16 processor cores, 2.2 GHz each. The robot is equipped with various sensors such as Firewire camera, laser range finder, sonar array, IMU, compass, and GPS. In addition, panel-mount waterproof USB connectors are available for new sensors, along with RJ45 Ethernet for wired Internet connection and panel-mount KVM inputs for regular-sized monitor, keyboard, and mouse. There is also a touchscreen LCD for a convenient user interface. Furthermore, the kill switches at each corner of the robot are available as a last resort to stop it in emergency situations.

depth below), and implementing signals such as gigabit Ethernet for the backplane intermodule communication, as well as others such as SATA (two per module), PCI Express, USB, and VGA. Beobot 2.0 uses a PCI Express 1394-Firewire card for a low-latency camera connection. One of the SATA ports was used for the primary hard drive and the other for external drives such as CD-ROM (useful for installing operating systems, for example). Giving each module its own hard drive obviates the need to pass around copies of stored data, such as large knowledge databases obtained during training.

There are six USB signal implementations per computer for a total of 48. Some of them are being used for sensors listed in Table I. Several of the USB connectors are panel mounted outside the robot for ease of connecting external devices using dust- and waterproof connectors (Figure 1). In addition, there are also USB connectors inside, on the baseboards (see Figure 3).

Furthermore, there is an onboard keyboard-video-mouse (KVM) switch to toggle between each of the eight computers. The KVM is an eight-to-two switch, eight computers to two display outputs. The display signal outputs can be either a regular-sized external monitor or to an onboard 8-in. touchscreen LCD with a full-color video graphic array (VGA) interface (Figure 1). Note that in practice we operate all computers from a single node using an SSH login session to conveniently run and monitor multiple programs simultaneously. The use of wired interface to the individual computers is usually limited to hardware, BIOS, and boot troubleshooting.

**Table I.** Sensors provided in Beobot 2.0.

| Item | Company name | Reference |
|---|---|---|
| Laser range finder | Hokuyo | Hokuyo Automatic Co., Ltd., 2009 |
| IMU | MicroStrain | MicroStrain, Inc., 2009 |
| Compass | PNI | PNI Sensor Corporation, 2009 |
| Sonars (7 units) | SensComp | SensComp, Inc., 2009 |
| GPS | US Global Sat | USGlobalSat, Inc., 2009 |

The objectives for selecting a computing platform appropriate for the robot are high computing power, compactness, and low energy consumption. To have close to maximum achievable speed, we concentrate on the X86 architectures rather than far less powerful CPU types such as ARM or Xscale. Within the X86 family, we select the mobile processor version rather than its desktop counterpart for energy efficiency, still being competitive in computing power. By the same token, in using the mobile CPU version, the corresponding embedded systems option can be selected

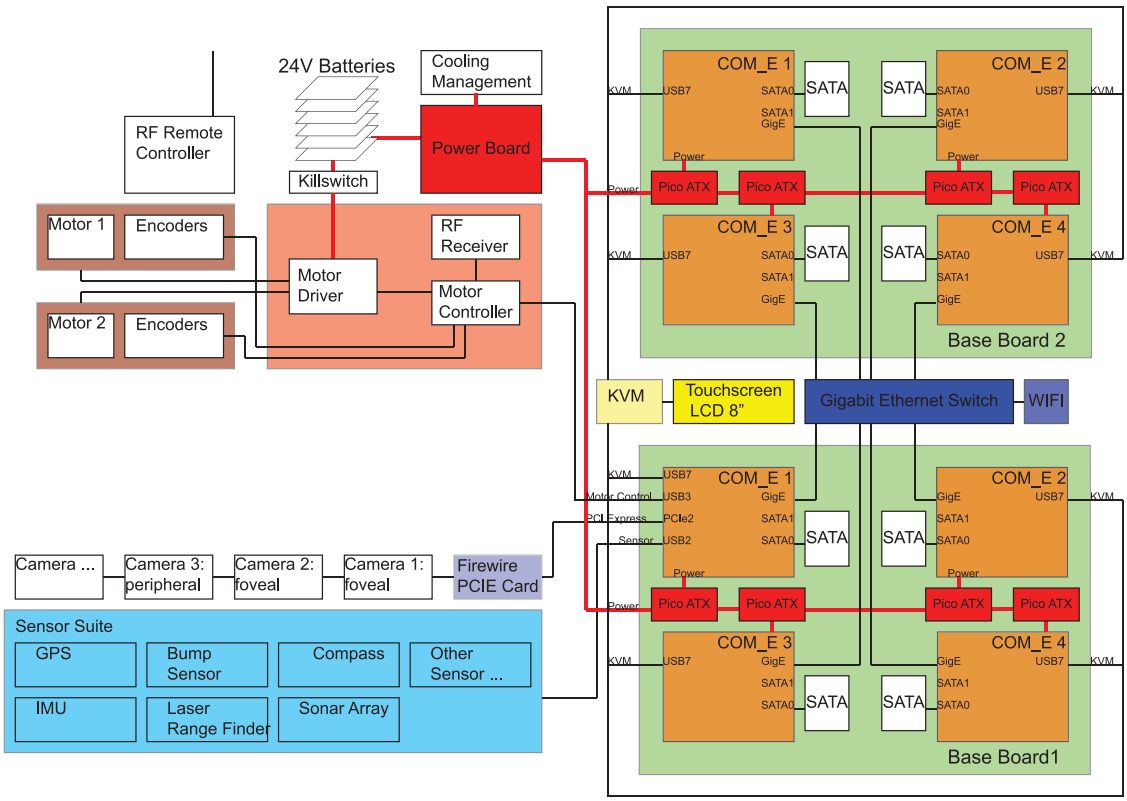



**Figure 2.** Beobot 2.0 electrical system. On the right-hand side of the diagram, there are two baseboards, each housing four COM express modules and each module with its own SATA hard drive. The backbone intercomputer communication is gigabit Ethernet that is connected through a switch. For visual interface to individual computers, a KVM is used to connect to either an 8-in. LCD touchscreen or an external monitor, mouse, and keyboard. In addition, a PCI Express–Firewire interface card is used to connect to a low-latency camera. The other sensors are connected via the many USB connectors that are panel mounted on top of the robot as well as on the baseboard. The whole system is powered by a 24-V battery circuit supply (with kill switches for safety purposes) and is regulated through a set of dedicated Pico-ATX power modules. The same battery circuit also powers the motors as well as the liquid-cooling system.

for the mobile platform (regular-sized motherboards do not usually use mobile CPUs), which resolves the size issue.

In the family of embedded systems, there are two types of implementations. The first family of systems have the interfaces already implemented, ready to use. An example is the ITX form-factor family (pico-ITX, nano-ITX, mini-ITX) (Via, 2009). The drawback is that the provided interfaces are fixed. They may not be the specific ones that are needed, and unused connections can be a waste of size as we cannot customize their location and orientation. In addition, by using off-the-shelf motherboards, their dimensions have to be accommodated in the design specifications, which may also limit the options for the locomotion platform.

In contrast, the second type of embedded systems, the COM concept, provides specifications for all the interfaces only through a set of high-density connectors. These specifications are usually defined by an industry consortium such as the XTX-standard (XTX Consortium, 2009). The actual breakout of the individual signals (such as gigabit Ethernet, USB, PCI Express) from the module connectors to the outside devices (a hard drive, for example) has to be done on a custom-made carrier board. By building custom baseboards, the overall size of the electronics can be controlled by implementing only those signals that we actually need. In addition, connector placement (as well as type) can be specified so as to minimize the amount of cabling in the system.

In the end, we found that a COM module solution best met our requirements, which we stated at the start of this section. Within this class, there are three options: ETX (ETX Industrial Group, 2009), XTX (XTX Consortium, 2009), or COM Express (COM Express Extension, 2009). These modules use the most powerful processors, as opposed to the smaller but less powerful systems such as PC104-based

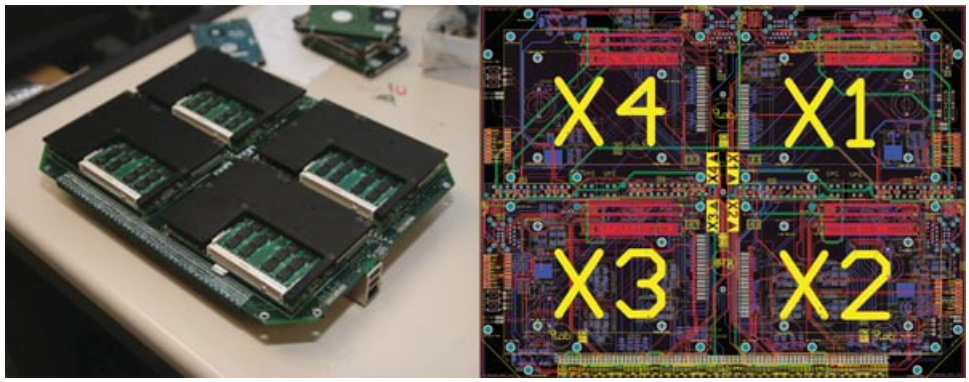



**Figure 3.** Baseboard. The image on the left is a fully assembled baseboard with four COM Express modules. The black plates are the heat spreaders attached to the processors. There is also an Ethernet and two USB jacks placed on the right-hand side of the board. The layout on the right is the circuit done in Altium (Altium Limited, 2009) PCB design software.

Qseven (Qseven Standard, 2009). We chose COM Express because it has an onboard gigabit Ethernet interface on the module, and it is only slightly larger (12.5-cm length × 9.5-cm width) than the XTX and ETX module (11.5-cm length × 9.5-cm width). Gigabit Ethernet is critical because in a cluster architecture, intercomputer communication can be just as important as the computing power of individual nodes. If the communication procedure cannot provide data fast enough, the individual processors will simply idle most of the time, waiting for data to arrive. This is especially true in our case because Beobot 2.0 is designed to perform heavy-duty, real-time vision computation in which the real-time video streaming is much more demanding than sending intermediate results.

We implemented two carrier/baseboards (refer to Figure 3), each accommodating four COM Express modules. A total of eight modules is chosen because the computing system fits within the mobile platform and because this number is expected to suffice for our research needs based on the findings presented in Section 5.

We used the Kontron COM Express design guide (Kontron, 2007) [for the Kontron ETX-Express-MC 2.2 GHz (T7500) COM Express module (Kontron, 2009)] to help properly design the electronic circuits as well as lay out the components in the board. We used the electronics computer-aided design (ECAD) layout software Altium (Altium Limited, 2009) to plan the physical placement of all the desired devices and connectors with as little cabling as possible for a system of eight computers. Altium's three-dimensional (3D) visualization proved to be an invaluable feature as it allowed us to verify that boards and devices packed close together in the robot would not collide or otherwise interfere with any other components.

The most critical part in successfully implementing the baseboards was being able to take care of the high-speed differential-pair signal requirements such as matching length and spacing, as well as minimizing the number of vias in the baseboard. Altium allows its users to specify rules for each trace on the board, which tremendously eases the process of identifying unsatisfied constraints. We found that the signals are quite robust as long as the stated requirements are adhered to. In addition, most of these signals need very few supporting circuits. The most components required by a signal is eight, for the USB current limiter (500 mA) circuit. The VGA signal actually specifies that it needs a filtering circuit with many components, but the KVM chip furnishes this feature.

To provide clean and fail-safe power given a supply from the available batteries, a Pico-ATX (Ituner Networks Corp., 2009) module [see Figure 4(a)] is used to regulate power to each COM Express for a total of eight. There is also one extra Pico-ATX powering all the peripheral boards

(a) Pico-ATX (b) KVM board (c) Liberty312 battery and charger

**Figure 4.** Various power-related components.

and sensors. There are three peripheral boards: one to control the cooling system, one to control access to the motors, and a sensor board that houses all the various built-in sensors. The power to the drive-train motors does not need to be filtered and, thus, is directly connected from the batteries. Because power supplies have a high rate of failure, going with multiple power modules provides better granularity in that if one module fails, the remaining seven computers can still run. Additionally, the lower individual supply requirement allows for a wider range of products to choose from than would have been available in a single module solution.

Table II summarizes the important electrical features: interprocessor communications, input/output interfaces, KVM interface, sensors, and power management. These features are shown to be critical while compiling the list of available commercial robots as well as from our experience conducting robotics research.

## 3. MECHANICAL SYSTEM DESIGN AND IMPLEMENTATION

The mechanical design of the robot is divided into two parts: the locomotion platform, which is the dark-colored robot base in Figure 5 and described in Section 3.1, and the computing cluster housing, which is the cardinal-colored structure described in Section 3.2.

Again, note that we created a wiki page (Siagian et al., 2009) to detail the execution matters such as actual part drawings (SolidWorks Corp., 2009), part manufacturing through a machine shop, or finding suppliers for the needed devices listed in the bill of material.

### 3.1. Locomotion System

For the locomotion platform, we selected a Liberty 312 electrical wheelchair (Major's Mobisist, 2009) instead of building one from scratch. Often priced at thousands of dollars, these types of units are easily acquired second hand through channels such as Craigslist or eBay (ours cost U.S. $200). The wheelchair is a robustly engineered, stable, safe, and low-maintenance platform. Most importantly, adhering to the wheelchair form factor allows the robot to traverse most terrain types encountered in modern urban environments, both indoors and out. This platform can also carry heavy payloads (113–136 kg), which means the ability to add many more devices to the robot's computing cluster. An important factor to consider is the ability to control the motors over a wide range of speeds (0–16.09 km/h) with good resolution in between. The wheelchair platform has this characteristic as it is designed for fine-grained control, as opposed to the remote control (RC) car used by the original Beobot (Chung et al., 2002), which could be driven only at maximum speed. Another benefit of the wheelchair is that it places the computing cluster on top, relatively high above the ground (about 50 cm) and away from the thick dust and mud that can accumulate on the street. Note that the robot's driving dynamics is taken care of because the wheelchair is designed to have a person on top, where the computing system now is placed. This is accomplished by the wide-spacing configuration of the wheels, enveloping the payload, to allow for the overall balance of the system while it is moving reasonably fast. In addition, the heavy SLA batteries are placed on the bottom to lower the center of mass.

To control the wheelchair, we designed a motor board to connect the battery and motors to inputs from the computer for autonomous control as well as to a 2.4-GHz remote controller (RC) for manual driving or overriding. A dual-output motor-driver named Sabertooth (Dimension Engineering LLC, 2009) is used to provide up to 25 A to each motor. In addition, because the motor driver has a built-in electrical brake system, the mechanical brakes that stop the motors by pinching the back shafts are taken off. This then allows the back shafts to be coupled to a pair of encoders to provide odometry data. As a safety precaution, Beobot 2.0 is furnished with four kill switches (Figure 1), one on each corner for the user to immediately stop the robot in the event of an emergency.

The wheelchair comes with a pair of 12-V, 35-Ah SLA batteries, connected in series to provide a 24-V supply. They have a form-factor space of 19.5-cm length $\times$ 13.2-cm width $\times$ 15.5-cm height for each battery. An attractive feature of the wheelchair is the built-in, wall-outlet, easy-plug-in battery recharging system, shown in Figure 4(c). With this, the batteries can be conveniently recharged without having to put them in and take them out of the robot, although the recharging process does take an average of 10 h.

### 3.2. Computing Cluster Case

The structure surrounding the computing clusters, as shown in Figure 5, shields the computing cluster from unwanted environmental interference such as dust and mud. The structure is divided into two isolated chambers as illustrated in the figure. The back chamber is the watertight area where the cluster is placed. The front chamber is an open area, reserved for a liquid-cooling system (further elaborated in Section 3.2.2), which includes a radiator to allow for maximum air flow. These two cooling subsystems are connected through Tygon tubing for liquid flow and are physically held together by a pair of aluminum holders. The computing cluster, along with the cooling system, itself is mounted on shock-absorbing standoffs (Section 3.2.1) to withstand violent collision in the rare event the robot hits an obstacle.

#### 3.2.1. Vibration Attenuation and Shock Absorption System

As illustrated in Figure 5, the only connections between the computing system and the robot base are the

**Table II.** Beobot 2.0 electrical system features.

| Feature | Our requirements | Solution chosen | Alternative considered | Positives of chosen solution | Negatives of chosen solution | Remarks |
|---|---|---|---|---|---|---|
| Intermodule communication | Large enough bandwidth to stream real-time video | Gigabit Ethernet | Symmetric multiprocessing (SMP) architecture, same memory module, shared throughout the system bus | Simpler to build | Larger latency | Gigabit Ethernet network also connected to wireless Internet connection for remote logins |
| Input/output interfaces | Connection to gigabit Ethernet, PCI Express, SATA, USB, and VGA signals from the COM Express modules | Custom mother/baseboards and peripheral boards connected to easily accessible panel-mount dust and waterproof connectors | Cables for access from COM Express modules to the connectors | Minimal cabling allow for easier debugging/repair | Complex design, especially in implementing high-speed signals with differential pair requirements | Found that the signals are quite robust as long as the stated requirements [in the design guide (Kontron, 2007)] are adhered to |
| KVM computer interfaces | Provide easy visualization and access to each computer | Integrated KVM system | Individual VGA, mouse, keyboard, cables | Do not need to cram cables to eight computers into a small area | Complex design, integrated in the baseboard and peripheral boards | 8 computers to 2 display KVM switches: one for regular-sized external monitor (if desired), another onboard, 7 in full-color VGA touchscreen LCD; note that software remote access solution (using SSH, for example) is also available |

(*Continued*)

**Table II.** Continued

| Feature | Our requirements | Solution chosen | Alternative considered | Positives of chosen solution | Negatives of chosen solution | Remarks |
|---|---|---|---|---|---|---|
| Sensors | Need sensors for a wide range of robotics and vision research | Integrate sensors to the design | Add sensors later | Less likely to accommodate space for sensors | Upfront added cost weight and design complexity | We include versatile sensors that are widely used; also has an abundance of accessible USB ports throughout robot body and microcontrollers for new sensors |
| Power Management | Clean and fail-safe power (12, 5, and 3.3 V) as per COM Express Design Guide (Kontron, 2007); these modules primarily need large supply of 12 V, up to 5-A maximum per module, making for a total of 40 A (480 W) for all eight modules | Off-the-shelf system: Pico-ATX (Ituner Networks Corp, 2009) modules to power each COM Express board (plus 1 for all the peripherals and sensors) | Built custom power system and profile supplying high current on the voltages needed the most | Fully engineered and tested, with protection from overdischarging and overpeaking; shorter prototyping cycles; end up with about the same cost | Need to accommodate shape and sizes, waste of energy for unneeded supplies (−5 and −12-V CC), and lack of availability of a 12-V supply that goes over 12 A | Modified the compact Pico-ATX and routed the power traces appropriately in the baseboard to minimize cable length; they fit in the available space without requiring any alterations to the mechanical design |

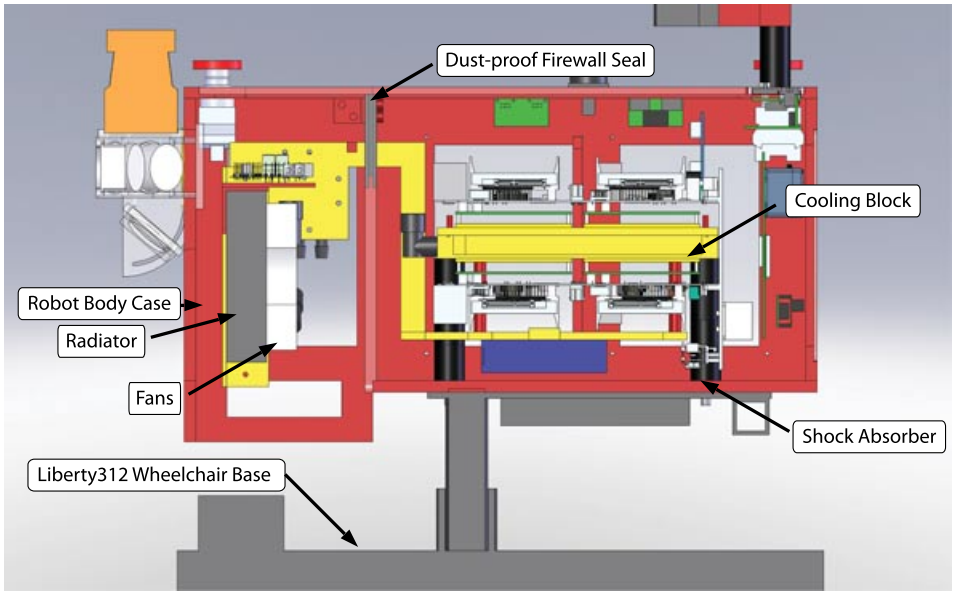



**Figure 5.** A SolidWorks (SolidWorks Corp., 2009) model of the robot shown from its side and cut through its center. The bottom of the image displays the Liberty312 wheelchair base, and above it is the robot body in cardinal color. The robot body is divided into two chambers by the dust-proof firewall. The back chamber completely seals the computers within from the elements. The front of the robot, which houses part of the cooling system, is open to allow for heat dissipation. The heat from the computers is transferred by the liquid in the cooling block, which is attached to the heat spreaders on each module. The liquid then moves through the radiator, which is cooled by the fans, before going back to the cooling block. In addition, the computing block is shock mounted on four cylindrical mounts that are used to absorb shocks and vibration.

shock-and-vibration damping standoffs. This makes it easier to properly evaluate the necessary damping requirements. When considering a damping solution, one needs to take into account the basic relationship between shock and vibration. That is, the solution has to be rigid enough to not cause too much vibration on the load but flexible enough to absorb shocks. Here, the focus is more on shock because, like regular laptops, the computers should be able to work despite the vibration that comes from reasonably rough terrains. In addition, the system uses solid-state hard drives (SSD), which have no moving parts and can withstand far more shock than their mechanical counterparts.

The natural rubber cylindrical mounts are selected over other options such as wire–rope isolators, rubber or silicone pads, and suspension springs because of their compactness. In addition, the height of the standoffs is easily adjustable by screwing together additional absorbers according to needs. Furthermore, one can change their shock absorption property by adding washers between two mounts if need be.

### 3.2.2. Cooling System

Because Beobot 2.0 is meant to be used both indoors and outdoors, we decided against an air-cooling system due to the possibility of the fans pushing dust into the exposed electronics inside, although air filters could have kept the dust out. However, the electronics would have to be placed in an area where air flow is well controlled, i.e., air must be drawn in and exhausted out only through the fans. This would have entailed a push-and-pull fan system and significant prototyping and rework of the mechanical system.

Therefore, we settled on a liquid-cooling solution. Moreover, as water has 30 times the amount of thermal conductivity and four times the amount of heat capacity as air (Callister, 2003), a liquid-cooling system is more effective in addition to being cleaner.

The liquid-cooling system, as shown in Figure 5, consists of the following components: cooling block, tubes, nozzles, radiator, two fans, liquid pump, reservoir, cooling liquid, a flowmeter, and a temperature sensor to monitor the system. Note that the system uses a cooling control board to provide power for the fans and the pump, as well as to take data from the flowmeter and temperature sensor.

The heat dissipated by the COM Express modules is first transferred to the liquid coolant through the processors' heat spreaders that are firmly pressed up against the top and bottom of the cooling block, which contains the coolant. We recommend using a high-performing, low-conducting, noncorrosive coolant for a maintenance-free system. The heat-carrying coolant first goes through the radiator, which has two fans pulling air through the radiator

surface. These fans are the devices that actively take the heat out of the system. Note that the radiator (and the fans) can be placed as far away from the processors as necessary. The liquid pump is connected to the system to ensure the flow of the liquid. Finally, a reservoir is included to add the coolant into the system and to take the air (bubbles) out of it.

## 4. SOFTWARE DESIGN

Our ultimate goal is to implement a fully autonomous embodied system with complete visual scene understanding. To do so, we lay the groundwork for a robot development environment (Kramer & Scheutz, 2002) that especially maximizes the multiple-processor hardware architecture. In addition it fulfills the primary objective in designing the software, viz., to be able to integrate and run computationally heavy algorithms as efficiently as possible. The advantage of using COM Express modules as a platform is that they can be treated as regular desktops. This allows the use of a Linux operating system in conjunction with C++ rather than some special-purpose environment. Note that, this way, the user can install any kind of Linux-compatible software tools that he/she prefers, not just the ones that we suggest below. Also, although this is not a true real-time system, it is quite adequate for our needs, with the control programs running reasonably fast and the robot responding in real time. In case a user would like to go with a real-time operating system, several Linux-based options and extensions are available (Politecnico di Milano, 2010; QNX Software Systems, 2010; Wind River, 2010; Xenomai, 2010).

To speed up the development of the complex algorithms mentioned above, we use the freely available iLab Neuromorphic Vision C++ Toolkit (Itti, 2009). The motivation for the toolkit is to facilitate the recent emergence of a new discipline, neuromorphic engineering, which challenges classical approaches to engineering and computer vision research. These new research efforts are based on algorithms and techniques inspired from and closely replicating the principles of information processing in biological nervous systems. Their applicability to engineering challenges is widespread and includes smart sensors, implanted electronic devices, autonomous visually guided robotics systems, prosthesis systems, and robust human–computer interfaces. Thus, the development of a neuromorphic vision toolkit helps provide a set of basic tools that can assist newcomers in the field with the development of new models and systems.

Because of its truly interdisciplinary nature, the toolkit is developed by researchers in psychology, experimental and computational neuroscience, artificial intelligence, electrical engineering, control theory, and signal and image processing. In addition, it aids in integration with other powerful, freely available software libraries such as Boost and OpenCV.

The project aims to develop next-generation general vision algorithms rather than being tied to specific environmental conditions or tasks. To this end, it provides a software foundation that can be used for the development of many neuromorphic models and systems in the form of a C++ library that includes classes for image acquisition, preprocessing, visual scene understanding, and embodied system control.

These systems can be deployed in a single machine or a distributed computing platform. We use the lightweight middleware ICE (Internet Communication Engine) via its C++ library bindings to facilitate intercomputer communication with a high-level interface that abstracts out low-level matters such as marshaling data and opening sockets. Sensors/devices, which are connected to a computer in the distributed system, are encapsulated as independent services that publish their data. Different systems can grab just the sensor outputs that they need by subscribing to that particular service. In addition, such a distributed system is fault tolerant as nonfunctional services do not bring down the whole system. We are also working on adding functionality to quickly detect nonresponding hardware and recover from failures by performing an ICE reconnection protocol, for example.

Another aspect to pay close attention to is the need for robust debugging tools for distributed systems that are provided by the toolkit as well as future applications. That is, we would like to know which modules in the system take the longest times, which ones send the largest amount of data, and how all these factors affect the overall system efficiency. Currently, the system has logging facilities for analysis after a testing run has taken place. What would be ideal is an online monitoring system.

In terms of hardware support, the toolkit has extensive source code available for interfacing sensors through different avenues. For example, Beobot 2.0 currently can connect to different types of cameras: USB, Firewire, or IP. Other devices that use a serial protocol should also be easily accommodated. In addition, it is important to note that the separation of hardware-related and algorithm-related code comes naturally. This allows the user to test most of the software in both the robot and our custom simulator (provided in the toolkit) without too many changes. Furthermore, the same robot cluster computing design is used for our robot underwater and aerial vehicles. We find that porting the algorithms to the other robots is done quite easily.

Table III lists all the vision- and robotic-related software capabilities provided by the toolkit.

## 5. TESTING AND RESULTS

We examine a few aspects of Beobot 2.0. The first is basic functionality such as power consumption, the cooling system, and mobility as it pertains to shock absorption. The power consumption testing shows the typical length

**Table III.** The vision toolkit features.

| Features | Description | Available options |
|---|---|---|
| Devices | Interface code for various devices | Embedded systems/microcontrollers, joystick, keyboard, gyroscope, wii-mote, GPS, IMU (HMR3300, MicroStrain 3DM GX2), LRF (Hokuyo) |
| Robots | Control code for various robots | Scorbot robot arm, Evolution Robotics Rovio, Irobot Roomba, Beobot, Beobot 2.0, BeoSub Submarine, BeoHawk Quadrotor aerial robots, Gumbot for undergraduate introduction to robotics |
| Robotics algorithm | Modular mobile robotics algorithm | Localization, laser, and vision navigation (lane following, obstacle avoidance), SLAM |
| Distributed programming tools | Allows programs to communicate between computers | CORBA, Beowulf, ICE |
| Neuromorphic vision algorithms | Biologically plausible vision algorithms | Center-surround feature maps, attention/saliency (multithreaded, fixed point/integer), gist, perceptual grouping, contour integration, border ownership, focus of expansion, motion |
| Media | Access to various input media | mpeg, jpeg, cameras (USB, IP, IEEE1394 Firewire), audiovisual |
| Image processing | Various tools to manipulate images | Drawing, cut/paste, color operations [hue saturation value (HSV), RGB, etc.], statistical operations, shape transformation, convolutions, Fourier transform, pyramid builder, linear algebra/ matrix operation |
| Machine learning | Tools for pattern recognition training | K-nearest-neighbor, backpropagation neural networks, support vector machine, genetic algorithm |
| Object recognition | Visual object recognition modules | SIFT, HMAX |

of operation given the amount of capacity of the batteries and the weight that the motors have to move and the eight computers that the batteries have to power. Beobot 2.0 has a power supply of 35 Ah $\times$ 24 V capacity from two 12-V SLA batteries in series. The robot is run with full-load computing by running a vision localization system (Siagian & Itti, 2009), explained in Section 5.2, while the robot is run around. In the testing, the cooling system is shown to drain about 1.8 A of the 24-V supply, whereas the gigabit switch and other sensors consume about 0.5 A. Each of the eight computers pulls up to 0.7 A during heavy use, and the motors pull 2 A when the robot is moving at about 1.61 km/h. The total comes up to 9.9 A in regular use, which corresponds to about 3.5 h of expected peak computation running time.

The good news is that Beobot 2.0 has two accessible power jacks located on its back, in the KVM board, as shown in Figure 4(b). By plugging in an auxilliary power source that stops its current flow when it detects another supply in the system, we can perform hot swapping to temporarily replace the SLA batteries. This prolongs the running time considerably, given that on-site system debugging occurs quite often. Consequently, the running time becomes actual testing time, without debugging time. This, for the most part, allows users to do research on site for the whole day and charge all night.

Table IV summarizes the results.

We then go into the usability of the system by reporting our experience implementing the nearness diagram (ND) navigation system (Minguez & Montano, 2004) in Section 5.1. Note that this section is included to show that the robot can move about an environment and is ready for use. We do not try to optimize the implementation to improve the performance. On the other hand, in Section 5.2, we describe our experiment performing three computationally intensive algorithms: the SIFT (Lowe, 2004) object recognition system, distributed visual saliency (Itti et al., 1998), and the robot vision localization system (Siagian & Itti, 2009). These computational speed/throughput experiments test the most critical aspect of the project's objectives. Given the complexity of having to implement a cluster of processors, we would like to see a good payoff for all our hard work.

### 5.1. Navigation Algorithm Implementation

In this section, we test the first algorithm to successfully run on Beobot 2.0, viz., the ND navigation algorithm (Minguez & Montano, 2004), which uses a laser range finder to build a proximity map around the robot and then searches this map for the navigable region closest to a goal location. A navigable region is a space or area that is at least as wide as the robot, thus enabling it to pass through. For example, the system's graphical user interface (GUI) display in Figure 6

**Table IV.** Beobot2.0 subsystem testings.

| Subsystem | Tests | Results | Remarks |
|---|---|---|---|
| Liquid cooling | CPUs at full load at room temperature of 22°C | Average CPU temperature of 41°C | System is virtually maintenance free, although it consumes 1.8 A for the liquid pump; CPUs reach critical overtemperature within 15 min if liquid-cooling system is turned off |
| Mobility and shock absorption | Running the robot throughout the campus using RF controller at 2 m/s | Computers run smoothly without disconnection through several bumps and abrupt stops whenever the robot is too close to nearby pedestrians | We modify the motor controller code to properly ramp down when going to a complete stop |
| Battery consumption | Running the robot using the remote controller with all computers running full computations | Robot runs for 2.25 h before one CPU shuts down | Can prolong the testing time considerably by hot swapping batteries (there is a jack at the back of robot) during on-site debugging; consequently, the running time becomes actual testing time, without debugging time, and allows us to do research on site for the whole day |

shows the robot's surroundings, divided into nine distinct regions.

The robot follows a series of binary decision rules that classify all situations into five mutually exclusive cases, which are summarized in Table V. Each case is associated with a corresponding movement action.

First, we define a security zone around the robot that is an area twice the robot's radius. In the GUI display (Figure 6), this zone is denoted by the light (yellow) center circle. If there are obstacles within the security zone (red dots within the circle in the figure), there are two cases to consider: whether there are obstacles on both sides of the robot or only on one side. In the former case, the robot tries to bisect this opening; in the latter case, it can move more freely to the open side. Note that the system considers only obstacles that are within 60 deg (between the two red lines in Figure 6) of the robot's direction of motion (blue line in the figure).

When there are no obstacles in the security zone, it considers three possible situations. If the goal location is in the navigable region, just go to it. If the goal location is not in the navigable region but the region is wide (only one obstacle on one of the sides), maneuver through the wide region, in the hope that there is a way to go to the goal region in the following time step. If goal location is not in the navigable region and the region is narrow (between two obstacles, one on each side), carefully move forward in the middle of the region. The overall resulting behavior is that the robot should continuously center itself between obstacles, while going to the goal.

To test this algorithm, only two of the available eight computers are needed. The laser range finder is plugged into one computer and the motor control board into the other. Additionally, a RC setup allows the user to change from autonomous to manual mode at the flick of a switch in case the robot is about to hit something or has been stuck in a corner for some time.

During implementation and debugging, a few notable features speed up the process. First, the 8-in. LCD screen allows users to observe the system states and action decisions as the robot is moving. Second, the use of a wireless USB keyboard and touch pad made it fairly easy to issue new commands while the robot was working. Last, but not least, taking the time to set up an intuitive GUI paid back dividends very quickly as it made it much easier to understand what was going on and how to fix the problems encountered.

The system is tested indoors, on a 20 × 24 ft empty area. We then occupy some of the regions with obstacles and test Beobot 2.0 to see whether it can navigate from one side of the environment and back. Figure 7 shows a snapshot of the environment setup for the experiment. In addition, some of the obstacle configurations are shown in Figure 8, with an example odometry trace overlaid on top.

There are nine different obstacle configurations and robot starting positions in the testing protocol. Each test was performed 10 times, with the robot's speed being the only variable parameter. We vary the speed between approximately 0.3 and 2.5 m/s. Table VI summarizes the results of each trial. For the most part, the navigation system

**Table V.** ND rules.

| Number | Situation | Description | Action |
|---|---|---|---|
| 1 | Low safety 1 | Only one side of obstacles in the security zone | Turn to the other side while maintaining the angle to the goal location |
| 2 | Low safety 2 | Both side of obstacles in the security zone | Try to center between both side of obstacles and maintain the angle to the goal location |
| 3 | High safety goal in region | All obstacles are far from the security zone and goal | Directly drive toward the goal |
| 4 | High safety wide in region | All obstacles are far from the security zone but goal is not in this region | Turn half max angle away from closest obstacles |
| 5 | High safety narrow in region | All obstacles are far from the security zone and narrower region in the goal location | Center both side of closest obstacles |

performs very well, with a 72% success rate. Here success is defined as the robot moving from its starting side of the environment to the other and back without touching any of the entities surrounding it.

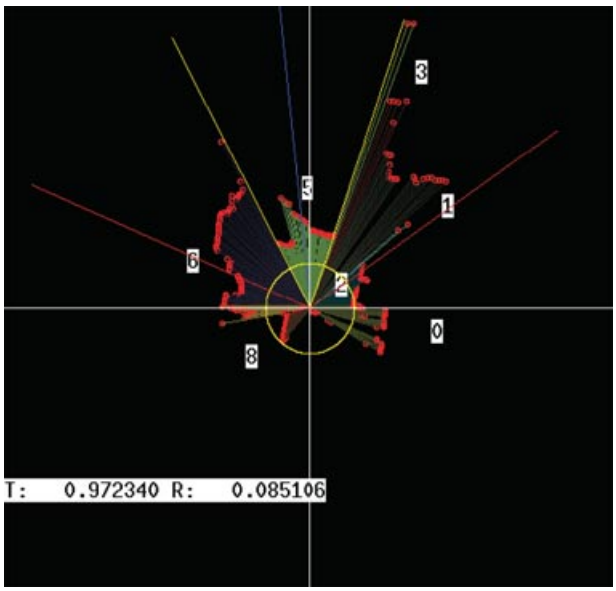



**Figure 6.** GUI of the ND navigation system. The system identifies nine (indexed from 0 to 8, note that 4 and 7 are cut off as they are drawn outside the frame of the GUI) different regions. The robot next direction of motion is denoted by the dark line next to label 5. The two lines next to the dark line delineate the boundaries of the navigable region. The red line indicates the directions 60 deg to the left and right of the robot's next direction. We also display the robot's translational and rotational motor command. Both of these numbers range from −1.0 to 1.0 (negative values indicate moving backward and counterclockwise rotation, respectively).

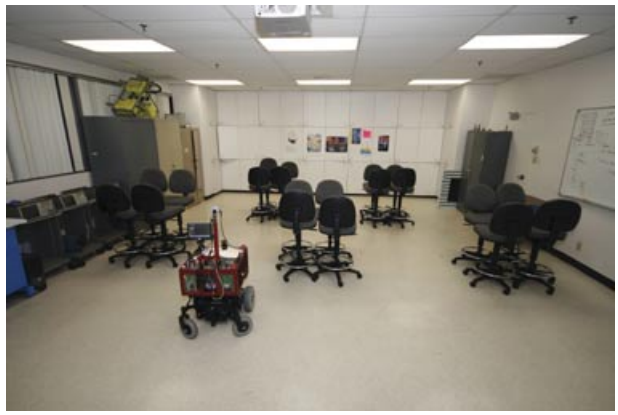
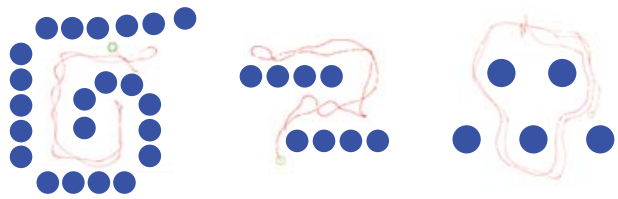


**Figure 7.** Snapshot of the constructed environment for ND navigation testing.



(a) Env1 (b) Env2 (c) Env3

**Figure 8.** Various environments for ND navigation testing with an example path that is taken by Beobot 2.0 using the ND navigation algorithm.

Of the total of 90 trials, 25 resulted in failures of some sort. Although this might seem excessive, it should be pointed out that the majority of the these collisions were of the type in which Beobot 2.0 only scraped an obstacle. This occurs whenever the robot has to turn sharply to avoid an obstacle, which causes its rear to scrape the obstacle. This is a minor problem that can be easily rectified by some simple

**Table VI.** Beobot 2.0 ND navigation testing.

| End result | Occurrence | Percentage |
|---|---|---|
| Success | 65 | 72.22 |
| Scraping the obstacles | 16 | 17.78 |
| Stuck in corner or circles | 7 | 7.78 |
| Squarely hit an obstacle | 2 | 2.22 |

control fix; e.g., when a turn is judged to be sharp, first back up a little.

There were two occasions when Beobot actually hit an obstacle head-on. This happened when the robot was running at its maximum speed. Under this circumstance, the latency of the system (the laser range finder throughput is 25 ms or 40 Hz, and processing is approximately the same duration as well) is simply too large to allow a timely reaction.

Finally, there were seven occasions when the robot became stuck in a corner or kept spinning in place because it kept alternating between left and right. The solution to this problem requires going beyond the simple reactive nature of the navigation system and figuring out what is globally optimal by integrating knowledge from localization or simultaneous localization and mapping (SLAM) algorithms.

## 5.2. Computational Capabilities

In this section, we characterize the computing platform by running three computationally intensive vision algorithms: SIFT (Lowe, 2004) object recognition system, the distributed visual saliency algorithm (Itti et al., 1998), and the biologically inspired robot vision localization algorithm (Siagian & Itti, 2009). These algorithms have a common characteristic in that their most time-consuming portions can be parallelized, whether it be distributing the feature extraction process (Section 5.2.2) or comparing those features to a large database (Sections 5.2.1 and 5.2.3). These parallel computations are then assigned to worker processes allocated at different computers in Beobot 2.0's cluster. Thus, we can fully test the computation and communication capabilities of the system.

### 5.2.1. SIFT Object Recognition System Test

As a first step in demonstrating the utility of our system in performing computationally intensive vision tasks, we implemented a simple keypoint matching system that is a very common component and performance bottleneck in many vision-based robotic systems (Se, Lowe, & Little, 2005; Valgren & Lilienthal, 2008). This task consists of detecting various interest points in an input image, computing a feature descriptor to represent each such point, and then searching a database of previously computed descriptors to determine whether any of the interest points in the current image have been previously observed. Once matches between newly observed and stored keypoints are found, the robot has a rough estimate of its surroundings and further processing such as recognizing its location or manipulating an object in front of it can commence.

Generally, the main speed bottleneck in this type of system is the matching between newly observed keypoints with the potentially very large database of previously observed keypoints. In the naive case this operation is $O(MN)$, where $M$ is the number of newly observed keypoints and $N$ is the number of keypoints in the database. However, this time can be cut to $O(M \log N)$ if the database of keypoints is stored as a KD-tree.

To test the efficacy of our cluster in speeding up such a task, we built a matching system in which a master node (called SIFT master) computes SIFT keypoints (Lowe, 2004) on an input image and then distributes these keypoints to a number of worker nodes (SIFT worker) for matching. Each of these workers is a separate process on the cluster located on either the same or a different machine as the master. Each worker contains a full copy of the database, stored as a KD-tree. Upon receiving a set of keypoints (different sets for each worker) from the master, a worker node compares each of them against its database and returns a set of unique IDs to the master representing the closest database match for each keypoint. Table VII describes the different types of modules in the system, and Figure 9 illustrates the flow of operation.

The database used for the experiment is composed of 905,968 SIFT keypoints obtained from HD footage (1,920 × 1,080 pixels) taken from an outdoor environment traversed by our robot. Each keypoint has 128 dimensions and eight bits per dimension. We vary the number of workers used to perform the keypoint matching from 1 to a maximum

**Table VII.** Beobot 2.0 SIFT object recognition time breakdown.

| Operation | Description | Computation time |
|---|---|---|
| Input SIFT keypoint extraction | Extract SIFT keypoints from the input image; done by the master process | About 18 s |
| SIFT database matching | Match the input keypoints with the SIFT keypoints and return the results; done by the worker processes | 16.25–353.22 s, depending on the number of workers utilized |

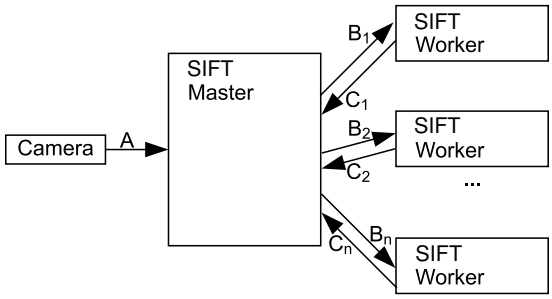



**Figure 9.** Flow of the distributed SIFT database matching algorithm denoted in increasing alphabetical order and referred below in parenthesis. First the camera passes in the high-definition (1,920 × 1,080 pixel) frame to the SIFT master module (A). This module takes about 18 s to extract the SIFT keypoints from the input image before sending them to the SIFT worker processes utilized (denoted as $B_i$, $i$ being the total number of workers). Depending on the number of workers, each takes between 16.25 and 353.22 s to return a match to the SIFT Master ($C_i$).

**Table VIII.** Beobot 2.0 SIFT database matching algorithm testing results.

| Number of workers | Processing time (s/frame) | Standard deviation (ms/frame) |
|---|---|---|
| 1 | 353.2194 | 57.8369 |
| 2 | 193.6876 | 58.1703 |
| 3 | 130.8932 | 39.6815 |
| 4 | 95.5375 | 27.8618 |
| 5 | 95.3156 | 34.4357 |
| 6 | 57.5809 | 12.2352 |
| 7 | 47.1917 | 10.6182 |
| 8 | 40.2749 | 9.9586 |
| 9 | 37.2474 | 8.3234 |
| 10 | 29.8436 | 6.6259 |
| 11 | 37.7632 | 15.9283 |
| 12 | 18.9525 | 6.2032 |
| 13 | 20.9672 | 6.1088 |
| 14 | 25.3063 | 6.8077 |
| 15 | 16.2541 | 5.7455 |

of 15 (because a total of 16 cores are available). Figure 10 illustrates the allocations of the modules. Table VIII and Figure 11 record the time required to process each frame, plotted against the number of workers.

Table VIII shows a total decrease of 21.73 times (from 353.22 to 16.25 s) in per frame processing time between 1 and 15 workers. Here, that the improvement goes beyond 15-folds is, we believe, because of memory paging issues that arise when dealing with the large messages necessary when using a small number of nodes.

Figure 11 also shows that while diminishing returns are achieved after 11 nodes, there is still a significant performance improvement by the utilization of parallel workers in the cluster. Although not all algorithms are as easily parallelized as this example, this experiment shows that a very common visual localization front end can indeed be parallelized and the benefits for doing so are significant.

#### 5.2.2. Distributed Visual Saliency Algorithm Test

One of the capabilities that is important in a robot is object collection. Here, a key task to perform is object recognition, usually from an image. There are times when the object may be small, or placed in a cluttered environment. This is when an algorithm such as the saliency model (Itti et al., 1998) can be quite useful. The term saliency is defined as a measure of conspicuity in an image, and by estimating this characteristic for every pixel in the image, parts of it that readily attract the viewer's attention can be detected. Thus, instead of blindly performing an exhaustive search throughout the input image, the saliency model can direct the robot to the most promising regions first. We can then equip the robot with a high-resolution camera to capture all the details of its surroundings. Furthermore, because of Beobot 2.0's powerful computing platform, saliency processing in such a large image in a timely manner becomes feasible.

To compute the salience of an image, the algorithm (Itti et al., 1998) first computes various raw visual cortex features that depict visual cues such as color, intensity, orientation (edges/corners), and flicker (temporal change in intensity). Here we have multiple subchannels for each domain: 2 color opponencies (red–green and blue–yellow center-surround computation), 1 intensity opponency (dark–bright), 12 orientation angles (increments of 15 deg), and 1 for flicker. That is a total of 16 subchannels, each producing a conspicuity map, which are then combined to create a single saliency map.

Because the computations in each subchannel are independent, they can be easily distributed. And so we use the algorithm to show how having many cores in a robot can alleviate such a large computational demand. For the experiment, we set up 1 master process and 1–15 worker processes to calculate the saliency of images of 4,000 × 4,000 pixels in size, for 100 frames. The master process takes approximately 100 ms to preprocess the input image before sending the jobs to the workers. The jobs themselves take up to 100 ms to finish for the color, intensity, and flicker subchannels and up to 300 ms for the 12 orientation subchannels. Finally, the conspicuity map recombination takes less than 10 ms. Table IX summarizes the running times of individual parts of the system, and Figure 12 illustrates the flow of the algorithm. In addition, Figure 13 shows the actual allocations of all the processes at which computer the modules are run.

The results that we obtained from this experiment can be viewed in Table X and are graphed in Figure 14. As we

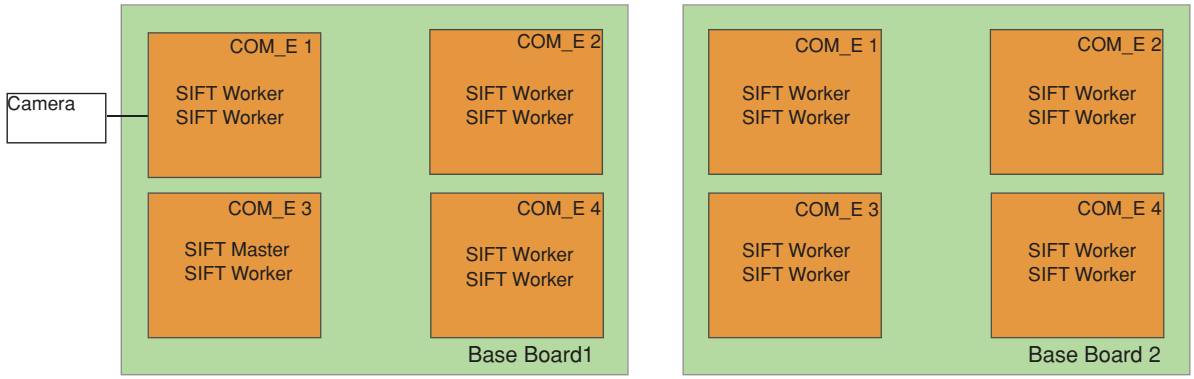



**Figure 10.** Allocation of the different programs of the distributed SIFT database matching algorithm in Beobot 2.0. The SIFT master module is run on one of the cores in computer COM_E1, and the various SIFT worker modules are allocated throughout the cluster.

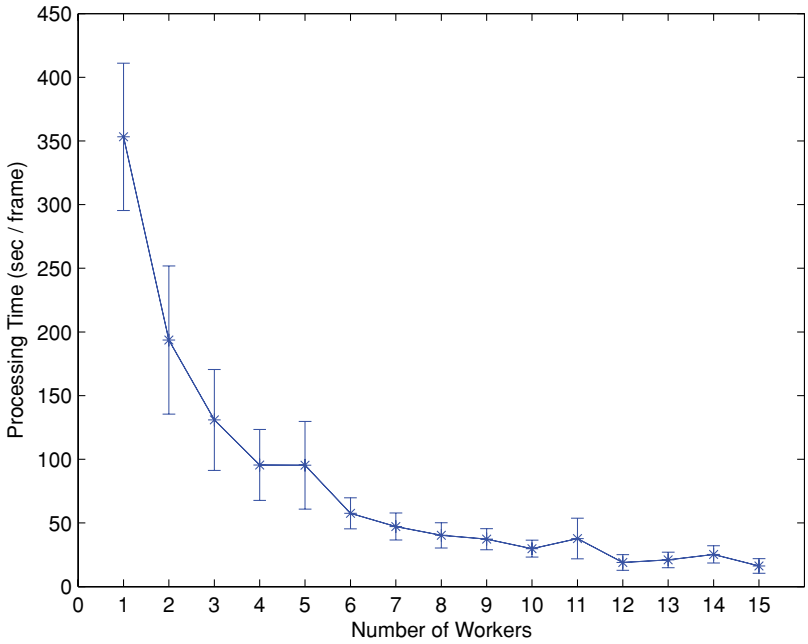



**Figure 11.** Results for SIFT database matching testing on Beobot 2.0.

can see, the processing time drops as we continue to add workers to the system. Quantitatively, the processing time reduction comes reasonably close to the expected value, at least early on. For example, if using one worker, the processing time is 3,456.50 ms, then using two workers should take half the time, 1,728.25 ms, which is comparable to the actual time of 1,787.69 ms. This is usually the case for a straightforward distributed processing in which there are no dependencies between the processes.

Another point of comparison is that we would like to gauge the improvement using the full cluster with what would be equivalent to a standard quad core system. Thus, we compare the usage of 3 worker nodes against all 15 nodes. We see a slight drop in improvement to 3.55 (from 1,249.68 to 352.26 ms). This slowdown is primarily attributed to network congestion, as we are shuffling large images around. Furthermore, if we compare the running time of 1 worker (3,456.5 ms) with 10 times the running time of 10 workers (478.33 ms $\times$ 10 $=$ 4,783.3 ms), there seems to be a lot of added time. And so, as we add more and more workers, we expect to eventually hit a point of diminishing returns. A lesson to be taken here is that we should consider not only how to divide the task and properly balance job allocation but also how large the data set (or the total communication cost) is that needs to be distributed for each assigned job.

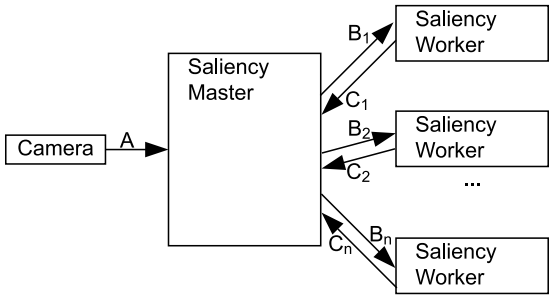



**Figure 12.** Flow of the distributed saliency algorithm denoted in increasing alphabetical order and referred below in parenthesis. First the camera passes in the high-resolution 4,000 × 4,000 pixel image to the SalMaster module (A). SalMaster preprocesses the image, which takes 100 ms, before sending out the image to various subchannel SaliencyWorker processes (denoted as $B_i$, $i$ being the total number of workers). The color, intensity, and flicker subchannels take up to 100 ms, and the orientation subchannels take up to 300 ms. These results are then recombined by SalMaster ($C_i$), and this takes less than 10 ms.

**Table IX.** Beobot 2.0 distributed saliency algorithm time breakdown.

| Module | Description | Computation time (ms) |
|---|---|---|
| Input image preprocessing | Computes luminance and red–green and blue–yellow color opponency maps to be sent to the worker processes; done by the master process | 100 |
| Conspicuity map generation | Performs center-surround operations in multiple scales to produce a conspicuity map for each subchannel | 300–3,900; depends on the number of workers utilized |
| Saliency map generation | Combines all the conspicuity maps returned by the workers to a single saliency map; done by the master process | 10 |

#### 5.2.3. Biologically Inspired Robot Vision Localization Algorithm Test

For the third computational test, we utilized the vision localization algorithm by Siagian and Itti (2009). It relies on matching localization cues from an input image with a large salient landmark database obtained from previous training runs to capture the scenes from the target environment under different lighting conditions.

The algorithm first computes the same raw visual cortex features that are utilized by the saliency algorithm (Itti et al., 1998). It then uses these raw features to extract gist information (Siagian & Itti, 2007), which approximates holistic aspects and the general layout of an image, to coarsely locate the robot in a general vicinity. In the next step, the system then uses the same raw features to isolate the most salient regions in the image and compare them with the salient regions stored in the landmark database to refine its whereabouts to a metric accuracy. The actual matching

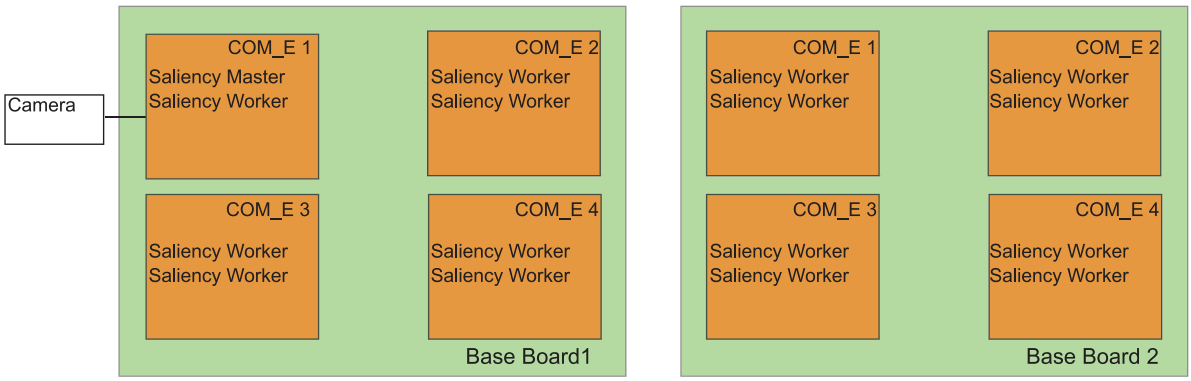



**Figure 13.** Allocation of the different programs of the distributed saliency algorithm in Beobot 2.0. The saliency master module is run on one of the cores in computer COM_E1, and the various saliency worker modules are allocated throughout the cluster.

**Table X.** Beobot 2.0 visual saliency algorithm testing results.

| Number of workers | Processing time (ms/frame) | Standard deviation (ms/frame) |
|---|---|---|
| 1 | 3,456.50 | 108.904 |
| 2 | 1,787.69 | 491.056 |
| 3 | 1,249.68 | 528.787 |
| 4 | 979.14 | 374.978 |
| 5 | 733.85 | 359.345 |
| 6 | 629.24 | 387.554 |
| 7 | 571.79 | 429.028 |
| 8 | 526.77 | 288.188 |
| 9 | 498.49 | 441.441 |
| 10 | 478.33 | 482.481 |
| 11 | 452.13 | 290.235 |
| 12 | 436.75 | 253.642 |
| 13 | 375.59 | 299.921 |
| 14 | 362.72 | 195.126 |
| 15 | 352.26 | 332.128 |

between two regions itself is done using SIFT features (Lowe, 2004). Of the different parts of the algorithm, the salient region recognition process takes the longest time. However, the computations performed by this module are parallelizeable by dispatching workers to compare particular parts of the database. Aside from the parallel searches, there are two other processes whose jobs are to extract gist and saliency features from the input image and a master process that assigns jobs and collects results from all landmark database search worker processes.

The gist and saliency extraction process, which operates on 160 × 120 size images, takes 30–40 ms to complete per frame and has to be run first. The images are placed in the computer that is connected to the camera. For this experiment, however, we are running off of previously collected data, without running the motors. Note that because the information being passed around consists of a few small salient regions (about five regions of 40 × 30 pixels, on average), only a small amount of time (4–5 ms) is spent on data transfer (using the ICE protocol) through the gigabit Ethernet network.

We then run the master search process, which takes about 50–150 ms (depending on the number of landmarks in the database) to create a priority queue for ordering landmark comparisons from most to least likely using saliency, gist, and temporal cues. For example, in the gist-based prioritization, if the gist features of the input image suggest that the robot location is more likely to be in a certain vicinity, we compare the incoming salient region with the stored regions found near that place. This prioritization improves the system speed because we are trying to find only a first match, which halts the search once it is found, and not the best match, which requires an exhaustive search through the entire database.

After these two processes are done, we can then dispatch the landmark search processes in parallel. For testing purposes, we use one, two, four, and eight computers to examine the increase in overall system speed. Noting that

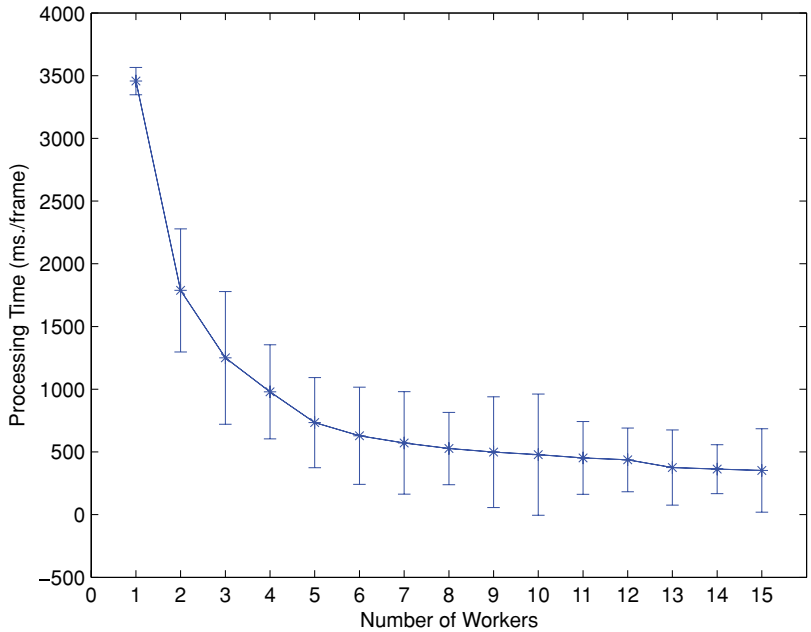



**Figure 14.** Results for saliency algorithm testing on Beobot 2.0.

**Table XI.** Beobot 2.0 localization system time breakdown.

| Module | Description | Computation time |
|---|---|---|
| Gist and saliency | Computes various raw visual cortex features (color, intensity, and orientation) from the input image for gist and saliency feature extraction | 30–40 ms |
| Localization master | Creates a priority job queue (to be sent to the workers) for ordering landmark database comparisons from most to least likely using saliency, gist and temporal cues; it also collects the search results from all search workers | 50–150 ms; depends on the size of the database |
| Localization worker | Compares the incoming salient region with the stored regions based on the prioritization order; the search halts once the first positive match is found | 300–3000 ms; depends on the size of the database and the number of workers utilized |

there are two cores in each computer, we dispatch 2, 4, 8, and 16 landmark database worker processes, respectively. The localization master then collects the match results to deduce the robot's most likely location given the visual evidence. This final step takes less than 5 ms.

Table XI summarizes the various processes, Figure 15 shows the program allocation, and Figure 16 illustrates the flow of the algorithm.

We test the system on the same data set as that used in Siagian & Itti (2009), which depicts a variety of visually challenging outdoor environments from a building complex (ACB) to parks full of trees (AnFpark) to an open area (FDFpark). The database information for each site in their respective rows, can be found in Table XII, and the images can be viewed in Figure 17. The table shows the number of training sessions, each of which depicts a different lighting condition in the outdoor environments. This is one of the reasons why the database is so large. The table also introduces the term salient region (Siagian & Itti, 2009), denoted as SRegs, which is different from a landmark. A landmark is a real entity that can be used as a localization cue, whereas a salient region is evidence obtained at a specific time. Thus there are, on average, about 20 salient regions to depict a landmark to cover different environmental conditions.

The results are shown in Table XIII. Here, we examine the processing time per frame, the localization error, and salient regions found per frame. As we can see from the table, for each site there is always a decrease in processing time per frame as we increase the number of computers. At the same time, generally, there is an increase in accuracy in two of the three sites as the number of computers is increased. The reason for this is that the localization

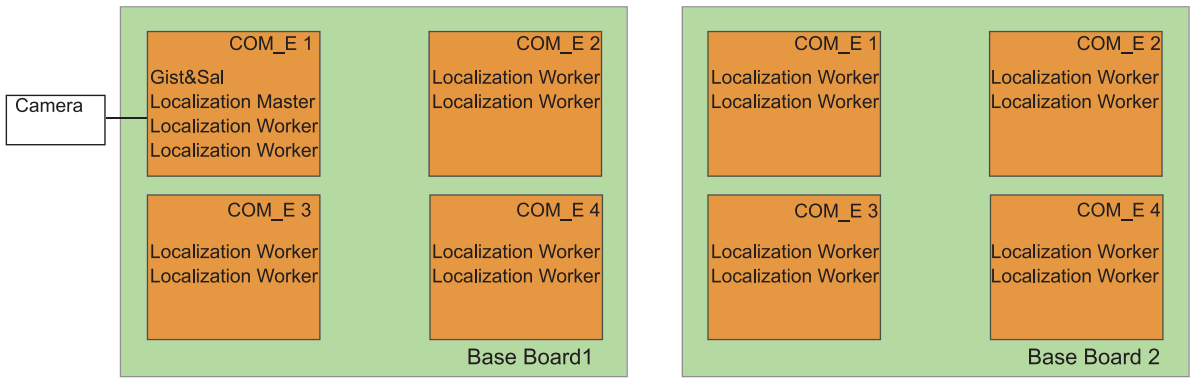



**Figure 15.** Allocation of the different programs of the localization system in Beobot 2.0. The gist and saliency extraction (GistSal) and localization master modules are allocated computer COM_E1, and the various localization worker modules are assigned to cores throughout the cluster. Note that there are also two worker modules in COM_E1. This is because they run only when GistSal and localization master modules do not, and vice versa.

**Table XII.** Beobot 2.0 vision localization system testing.

| Environment | Number of training sessions | Number of testing frames | Number of lmk | Number of S. Regs | Number of S. Reg/Lmk |
|---|---|---|---|---|---|
| ACB | 9 | 3,583 | 1,501 | 19.79 | 29,710 |
| AnFpark | 10 | 6,006 | 4,664 | 17.69 | 82,502 |
| FDFpark | 11 | 8,823 | 4,808 | 18.86 | 90,660 |

**Table XIII.** Beobot 2.0 vision localization system testing results.

| | ACB | | | AnF | | | FDF | | |
|---|---|---|---|---|---|---|---|---|---|
| Number of computers | Time | Err (m) | S. Reg found/frame | Time | Err (m) | S. Reg found/frame | Time | Err (m) | S. Reg found/frame |
| 1 | 1,181.60 | 2.21 | 2.34/4.89 | 2,387.87 | 2.27 | 2.73/4.98 | 3,164.96 | 4.30 | 2.51/4.78 |
| 2 | 711.93 | 1.70 | 2.38/4.89 | 1,495.23 | 2.31 | 2.76/4.98 | 1,909.01 | 4.36 | 2.51/4.78 |
| 4 | 499.18 | 1.13 | 2.48/4.89 | 1,000.66 | 2.36 | 2.81/4.98 | 1,201.90 | 4.04 | 2.55/4.78 |
| 8 | 421.45 | 1.26 | 2.57/4.89 | 794.31 | 2.38 | 2.94/4.98 | 884.74 | 4.08 | 2.60/4.78 |

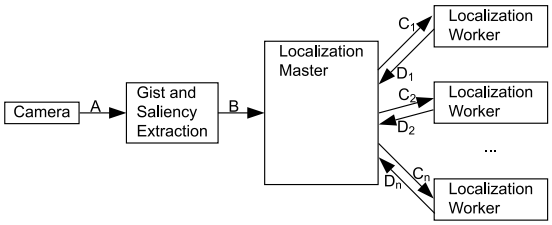



**Figure 16.** Flow of the localization system denoted in increasing alphabetical order and referred below in parenthesis. First the camera passes in a 160 × 120 pixel image to the gist and saliency extraction module (A), which takes 30–40 ms, before sending out the localization master module (B). This module then allocates search commands/jobs in a form of priority queue to be sent to a number of localization workers ($C_i$, $i$ being the total number of workers) to perform the landmark database matching. A search command job specifies which input salient region is to be compared to which database entry. This takes 50–150 ms, depending on the size of the database. The results are then sent back to the localization master ($D_i$) to make the determination of the robot location given the visual matches. The last steps takes less than 10 ms.

algorithm itself behaves differently as more and more resources are provided, in that it tries to optimize between the speed of computation and the accuracy of the results. Consequently, the running time analysis is not as straightforward. That is, we cannot just look at the nonlinearity of the relationship between the number of computers and processing time, stating doubling the number of computers does not halve the processing time, and say that the algorithm does not take advantage of the available computing efficiently.

As we explained earlier, the Siagian and Itti (2009) localization system orders the database landmarks from the most to the least likely to be matched. This is done by using other contextual cues (such as gist features, salient feature vectors, and temporal information) that can be computed much quicker than in the actual database matching process. The effect of this step is that it gives robot systems with limited computing resources the best possible chance to match the incoming salient regions. In addition, there is also an early-exit strategy that halts the search if the following conditions are met:

- Three regions are matched.
- Two regions are matched and 5% of the queue has been processed since the last match.
- One region is matched and 10% of the queue has been processed since the last match.
- No regions are matched and 30% of the queue has been processed.

This policy is designed to minimize the amount of unnecessary work when it is obvious that a subsequent match is very unlikely to be found. However, together with the increase in the number of workers, this policy actually creates a slightly different behavior.

First, there is a difference between the number of jobs processed by a one-worker setup compared to a multiple-worker setup. In the former setup, the localizer master process assigns a job, waits until the worker is done, and then checks whether any of the early-exit conditions are met before assigning another job. In the multiple-worker case, the master assigns many jobs at the same time and much more frequently. This increases the possibility of a match,

**Figure 17.** Examples of images in the ACB (first row), AnFpark (second row), and FDFpark (third row).

as demonstrated by the increase in the number of salient regions found in Table XIII. In turn, this slows the running time by prolonging the search process by 5%, 10%, or even 15% (in a compound case) of the queue, depending on which early-exit conditions are invalidated.

On the other hand, however, the higher number of matches found can also increase the accuracy of the system, but not always. As we can see in Table XIII, there is a small but visible adverse effect of letting the search go too long (most clearly in the AnF site). This is because the longer the search process, the more likely that a false-positive is discovered as the jobs lower in the priority queue are in lesser agreement with other contextual information. Furthermore, this is also reflected by the fact that a lot of the salient regions are found early in the search as the numbers do not increase significantly as we add more computers. For example, in the ACB site, compare the salient region found using one computer (2.34) with using eight (2.57).

From the table, we estimate that, for these environments, four computers appears to be the optimum number. Note that the localization system does not have to be real time, but being able to come up with a solution within seconds, as opposed to a minute, is essential because longer durations would require the robot to stop its motor and stay in place. This is what we are able to do with Beobot 2.0. In the full setup, the localization system is going to be run in conjunction with a salient region tracking mechanism, which keeps track of the regions while it is being compared with the database while still allowing the robot to move freely as long as the region is still in the field of view. If we use just four of the computers for localization, the others can be used for navigational tasks such as lane finding, obstacle avoidance, and intersection recognition, thus making the overall mobile robotic system real time. Currently, we have a preliminary result of a system that localizes and navigates autonomously in both indoor and outdoor environments, reported in Chang, Siagian, and Itti (2010).

## 6. DISCUSSION AND CONCLUSIONS

In this paper, we have described the design and implementation of an affordable research-level mobile robot platform equipped with a computing cluster containing eight dual-core processors for a total of 16 2.2-GHz CPUs. With such a powerful platform, we can create highly capable robotic applications that integrate many complex algorithms, use many different advanced libraries, and utilize large databases to recall information. In addition, by using the COM Express form-factor industry standard, the robot is able to stave off obsolescence for a longer period due to the ability to switch COM modules and upgrade to the latest processor and computer technology.

Furthermore, by implementing our own robot, we have demonstrated a cost-effective way to build such a computationally powerful robot. For more information on the cost breakdown, please refer to Siagian et al. (2009). The trade-off, of course, is in development time and effort. In our lab, we have had two people working full time on this project, with a few others helping out here and there. The total design and implementation time has been 18 months from conception to realization. We have had to think about many issues, no matter how small or seemingly trivial, in order to ensure that no major flaws are introduced into the design that can become showstoppers down the road. However, given that we now have the final design (Siagian et al., 2009), the implementation of a second robot ought to be relatively straightforward and much quicker, on the order of 2–3 months.

One might wonder why we would go to such as extraordinary effort to build such a complex hardware system when there may well be an easier alternative. For example, why not simply stack eight laptops on a mobile platform connected with Ethernet cables? We believe that with such an approach, it would be hard to isolate the computers from the elements. Cooling, for instance, would have to be done in two steps. First, the internal laptop fans would blow hot air out to a waterproof inner compartment of the

robot body. Then, we would need another set of fans, fitted with filters to drive the air in and out of the robot body. Furthermore, we would still have to create space to place all the other devices (for example, sensors and motor driver), along with power connections that also have to supply the main computers. The resulting shear numbers of cables would easily make the system unwieldy and unappealing.

In our custom design, however, wires and cabling, which are often a source of connection failures, have been kept to a minimum thanks to the printed circuit board (PCB) design directives. Additionally, the liquid-cooling system is well sealed and runs smoothly every time we run the robot. In terms of maintenance, we use the robot every day and have found it to be quite trouble free. The wall plug-in battery charging system, for one, makes it convenient to charge the robot at night before going home. Finally, we would like to add that, although in this paper we are presenting a terrestrial system, the same technology has been applied to an underwater robot (USC Robotics, 2009), where dimensions and weights become critical factors and modifying COTS (commercial off the shelf) systems may not be feasible.

Nonetheless, with the benefit of hindsight, there are some things we would have liked to improve upon. One is easier access to various electronic components inside the robot body. For example, four of the COM Express modules are placed underneath the cooling block structure, and taking them out for repair can be somewhat difficult. This is the price we pay for designing such a highly integrated system. Another problem is managing many computers. In an application that requires all eight of Beobot 2.0's computers, we have to compile and run many programs in parallel with certain ordering constraints. In addition, we also have to properly allocate where each program should be run on which computers, so that there are no computers that are idling while others are overloaded. Although these issues cannot always be avoided, some forethought and automation via appropriate scripting can help. Frameworks such as MOSIX or the Scyld Beowulf system are available to aid this process, which is to be tested in the future on our robot.

In the end, we believe that our primary contribution is that Beobot 2.0 allows for a class of computationally intensive algorithms that may need to access large-sized knowledge databases, operating in large-scale outdoor environments, something that previously may not have been feasible on commercially available robots. In addition, it also enables researchers to create systems that run several of these complex modules simultaneously, which is exactly what we are currently working on in our lab. That is, we would like to run the localization system (Siagian & Itti, 2009), vision-based obstacle avoidance, and lane following (Ackerman & Itti, 2005) together. We also are planning to add components such as SLAM and human/robot interaction. The long-term goal is to make available plenty of predefined robotic components that can be reused to speed up future project developments.

Subsequently, the problem that we foresee is managing these diverse capabilities. We have to make sure that there are enough resources to work with and give priority to the most important and reliable subsystems in solving the task at hand as well as identifying dangers that threaten the livelihood of the robot. We hope that through our contribution of implementing an economical but powerful robot platform, we can start to see more of the type of complete systems that are needed to thrive in a real-world environment.

## ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (grant number W911NF-08-1-0360). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

## REFERENCES

AAI Canada, Inc. (2009). AAI Canada, Inc.—Intelligent robots—Khepera II. http://www.aai.ca/robots/khep_2.html. Accessed October 15, 2009.

Ackerman, C., & Itti, L. (2005). Robot steering with spectral image information. IEEE Transactions on Robotics, 21(2), 247–251.

Altium Limited. (2009). Altium—Next generation electronics design. http://www.altium.com. Accessed July 15, 2009.

American Honda Motor Co., Inc. (2009). Asimo—The world's most advanced humanoid robot. http://asimo.honda.com/. Accessed July 15, 2009.

Bay, H., Tuytelaars, T., & Gool, L. V. (2006, May). SURF: Speeded up robust features. In ECCV, Graz, Austria (pp. 404–417).

Brooks, R. A. (1986). A robust layered control system for a mobile robot. IEEE Transactions on Robotics and Automation, 2(1), 14–23.

Callister, W. (2003). Materials science and engineering—An introduction. Hoboken, NJ: Wiley.

Carnegie Mellon University Robotics Institute. (2009). LAGR robot platform. http://www.rec.ri.cmu.edu/projects/lagr/description/index.htm. Accessed July 15, 2009.

Chang, C.-K., Siagian, C., & Itti, L. (2010, October). Mobile robot vision navigation & localization using gist and saliency. In Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan.

Chung, D., Hirata, R., Mundhenk, T. N., Ng, J., Peters, R. J., Pichon, E., Tsui, A., Ventrice, T., Walther, D., Williams, P.,

& Itti, L. (2002, November). A new robotics platform for neuromorphic vision: Beobots. In Lecture Notes in Computer Science (vol. 2525, pp. 558–566).

COM Express Extension. (2009). COM Express Extension—Consortium. http://www.comexpress-extension.org/specs/specs.php. Accessed October 15, 2009.

Dimension Engineering LLC. (2009). Sabertooth 2X25 regenerative dual motor driver. http://www.dimensionengineering.com/Sabertooth2X25.htm. Accessed July 15, 2009.

ETX Industrial Group. (2009). ETX Industrial Group—Consortium. http://www.etx-ig.org/consortium/consortium.php. Accessed July 15, 2009.

Evolution Robotics, Inc. (2009). Evolution robotics development platform and OEM solutions for robot software navigation technology, object recognition. http://www.evolution.com. Accessed July 15, 2009.

Finio, B., Eum, B., Oland, C., & Wood, R. J. (2009, October). Asymmetric flapping for a robotic fly using a hybrid power control actuator. In IROS, St. Louis, MO.

Fox, D., Burgard, W., Dellaert, F., & Thrun, S. (1999, July). Monte Carlo localization: Efficient position estimation for mobile robots. In Proceedings of Sixteenth National Conference on Artificial Intelligence (AAAI'99), Orlando, FL.

Fox, D., Burgard, W., Kruppa, H., & Thrun, S. (2000). A probabilistic approach to collaborative multi-robot localization. AutonomousRobots, 8(3), 325–344.

He, R., Prentice, S., & Roy, N. (2008, May). Planning in information space for a quadrotor helicopter in GPS-denied environments. In ICRA2008, Pasadena, CA.

Heitz, G., Gould, S., Saxena, A., & Koller, D. (2008, December). Cascaded classification models: Combining models for holistic scene understanding. In Advances in Neural Information Processing Systems (NIPS 2008), Vancouver, BC, Canada.

Hokuyo Automatic Co., Ltd. (2009). Photo sensor—PRODUCTS. http://www.hokuyo-aut.jp/02sensor/index.html#scanner. Accessed July 15, 2009.

iRobot Corporation. (2009a). PackBot. http://www.irobot.com/sp.cfm?pageid=171. Accessed July 15, 2009.

iRobot Corporation. (2009b). Roomba vacuum cleaning robot. http://store.irobot.com / category / index.jsp?categoryId=3334619&cp=2804605. Accessed July 15, 2009.

iRobot Corporation. (2010). SeaGlider. http://www.irobot.com/sp.cfm?pageid=393. Accessed January 15, 2010.

Itti, L. (2009). iLab Neuromorphic Vision C++ Toolkit (iNVT). http://ilab.usc.edu/toolkit/. Accessed July 15, 2009.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259.

Ituner Networks Corp. (2009). PicoPSU-80-WI-32V power supply. http://www.mini-box.com/PicoPSU-80-WI-32V?sc=8&category=981. Accessed July 15, 2009.

Kontron. (2007). Design guide for ETXexpress carrier boards. Poway, CA: Kontron.

Kontron. (2009). ETXexpress-MC. http://us.kontron.com / products / computeronmodules / com+express/etxexpress/etxexpressmc.html. Accessed July 15, 2009.

Kramer, J., & Scheutz, M. (2002). Development environments for autonomous mobile robots: A survey. Autonomous Robots, 22(2), 101–132.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

Maes, P., & Brooks, R. A. (1990, July–August). Learning to coordinate behaviors. In AAAI, Boston, MA (pp. 796–802).

Major's Mobisist. (2009). Major's Mobisist :: Power Wheelchairs :: Liberty 312. http://www.movingwithdignity.com/product.php?productid=16133. Accessed July 15, 2009.

MicroStrain, Inc. (2009). 3DM-GX2:: MicroStrain, AHRS Orientation Sensor. http://www.microstrain.com/3dm-gx2.aspx. Accessed July 15, 2009.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10), 1615–1630.

Minguez, J., & Montano, L. (2004). Nearness diagram (ND) navigation: Collision avoidance in troublesome scenario. IEEE Transactions on Robotics and Automation, 20(1), 45–59.

MobileRobots, Inc. (2009). Seekur unmanned ground vehicle. http://www.mobilerobots.com/Seekur.html. Accessed July 15, 2009.

MobileRobots, Inc. (2009). The High Performance All-Terrain Robot. http://www.activrobots.com/ROBOTS/p2at.html. Accessed July 15, 2009.

National Institute of Standards and Technology. (2009). Why are there no volume Li-ion battery manufacturers in the United States? http://www.atp.nist.gov/eao/wp05-01/append-4.htm. Accessed July 15, 2009.

PNI Sensor Corporation. (2009). PNI Sensor Corporation—Sensors modules—All products—MicroMag 3 : 3-Axis magnetometer. http://www.pnicorp.com/products/all/micromag-3. Accessed July 15, 2009.

Politecnico di Milano. (2010). RTAI—the RealTime Application Interface for Linux from DIAPM. https://www.rtai.org/. Accessed July 15, 2010.

Pomerleau, D. (1993). Knowledge-based training of artificial neural networks for autonomous robot driving. Robot learning (pp. 19–43). New York: Springer.

QNX Software Systems. (2010). QNX realtime RTOS—Middleware, development tools, realtime operating system software and services for superior embedded design. http://www.qnx.com/. Accessed July 15, 2010.

Qseven Standard. (2009). Qseven: About Qseven. http://www.qseven-standard.org/. Accessed October 15, 2009.

Quigley, M., & Ng, A. (2007, July). Stair: Hardware and software architecture. In AAAI 2007 Robotics Workshop, Vancouver, BC, Canada.

Salichs, M. A., Barber, R., Khamis, A. M., Malfaz, M., Gorostiza, J. F., Pacheco, R., Rivas, R., Corrales, A., Delgado, E., & Garca, D. (2006, June). Maggie: A robotic

platform for human–robot social interaction. In International Conference on Robotics, Automation, and Mechatronics (RAM 2006), Bangkok Thailand.

Se, S., Lowe, D. G., & Little, J. J. (2005). Vision-based global localization and mapping for mobile robots. IEEE Transactions on Robotics, 21(3), 364–375.

Segway, Inc. (2009). Segway—Business—Products & solutions—Robotic mobility platform (RMP). http://www.segway.com/business/products-solutions/robotic-mobility-platform.php. Accessed July 15, 2009.

SensComp, Inc. (2009). Mini Sensors. http://www.senscomp.com/minis.htm. Accessed July 15, 2009.

Siagian, C., Chang, C. K., Voorhies, R., & Itti, L. (2009). Beobot 2.0. http://ilab.usc.edu/wiki/index.php/Beobot_2.0. Accessed July 15, 2009.

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(2), 300–312.

Siagian, C., & Itti, L. (2009). Biologically inspired mobile robot vision localization. IEEE Transactions on Robotics, 25(4), 861–873.

SolidWorks Corp. (2009). SolidWorks 3D CAD Design Software. http://www.solidworks.com/. Accessed July 15, 2009.

Sony Entertainment Robot Europe. (2009). Aibo. http://support.sony-europe.com/aibo/. Accessed January 15, 2009.

Thrun, S., Fox, D., & Burgard, W. (1998). A probabilistic approach to concurrent mapping and localization for mobile robots. Machine Learning, 31, 29–53.

Thrun, S., Fox, D., Burgard, W., & Dellaert, F. (2000). Robust Monte-Carlo localization for mobile robots. Artificial Intelligence, 128(1–2), 99–141.

Thrun, S., & Liu, Y. (2003, October). Multi-robot SLAM with sparse extended information filters. In 11th International Symposium of Robotics Research, Siena, Italy (vol. 15, pp. 254–266).

USC Robotics. (2009). The SeaBee Autonomous Underwater Vehicle. http://ilab.usc.edu/uscr/index.html. Accessed July 15, 2009.

USGlobalSat, Inc. (2009). USGlobalSat Incorporated. http://www.usglobalsat.com/p-47-em-408-sirf-iii.aspx. Accessed July 15, 2009.

Valgren, C., & Lilienthal, A. J. (2008, May). Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In ICRA2008, Pasadena, CA.

Via. (2009). Pico-ITX Mainboard Form Factor. http://www.via.com.tw/en/initiatives/spearhead/pico-itx. Accessed February 15, 2009.

Willow Garage. (2009). PR-2—Wiki. http://pr.willowgarage.com/wiki/PR-2. Accessed July 15, 2009.

Wind River. (2010). Wind River: RTLinuxFree. http://www.rtlinuxfree.com/. Accessed July 15, 2010.

Wu, B., & Nevatia, R. (2007). Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. International Journal of Computer Vision, 75(2), 247–266.

Xenomai. (2010). Xenomai: Real-time framework for Linux. http://www.xenomai.org/index.php/Main_Page. Accessed July 15, 2010.

XTX Consortium. (2009). XTX: About XTX. http://www.xtx-standard.org/. Accessed July 15, 2009.