FINAL REPORT: April 1, 1996 - September 31, 1999

SPATIAL GRAPHICAL MODELS FOR IMAGE MATCHING AND OBJECT RECOGNITION

ARO Proposal no. 30167-MA
Grant no. - DAAH04-96-1-0061

Yali Amit
Department of Statistics
University of Chicago

February 18, 2000

FINAL PROGRESS REPORT

1 Summary

In this project we have had two partial successes. First is an efficient detection algorithm for objects in complex scenes, using very simple spatial arrangements to represent the objects, based on local features which are automatically identified in training. The simplicity of the arrangement allows us to use the Hough transform to very quickly find a small number of candidate locations for the objects. We have also proposed a parallel architecture for implementing this algorithm with interesting biological analogies. Second is an algorithm for isolated object recognition using decision trees to gradually explore the natural partial ordering of the space of spatial arrangements. The principles of this algorithm have also been successfully applied to the recognition of acoustic signals.

1.1 Shape recognition

A shape recognition algorithm has been developed based on multiple randomized decision trees. The splits in the trees are "queries" regarding the presence of partially invariant spatial arrangements of local features, *anywhere in the image.* These arrangements are defined through pairwise geometric relations between the features, and can be viewed as labeled graphs. As such they can be arranged in a partial ordering. Trees are grown by gradually exploring this partial ordering. All data images at a given node have one or more instances of the same arrangement present; the candidate splits entertained at that node are only minimal extensions of the arrangement allowing only one additional local feature and an additional relation.

Multiple trees are grown by choosing the best split from among a small random sample of all minimal extensions. The terminal distributions (conditional distribution on class at each node) of the trees are estimated from training data. A test point is classified by averaging

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	2.18.2000	Final 4.1.1996-9.31.1999

4. TITLE AND SUBTITLE
Spatial Graphical Models for Image Matching and Object Recognition

5. FUNDING NUMBERS
DAAH04-96-1-0061

6. AUTHOR(S)
Yali Amit

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Chicago

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. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 34172.3-MA

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

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

In this project we have had two partial successes. First is an efficient detection algorithm for objects in complex scenes, using very simple spatial arrangements to represent the objects, based on local features which are automatically identified in training. The simplicity of the arrangement allows us to use the Hough transform to very quickly find a small number of candidate locations for the objects. We have also proposed a parallel architecture for implementing this algorithm with interesting biological analogies. Second is an algorithm for isolated object recognition using decision trees to gradually explore the natural partial ordering of the space of spatial arrangements. The principles of this algorithm have also been successfully applied to the recognition of acoustic signals.

14. SUBJECT TERMS
Spatial Graphical Modes.
Object recognition and detection
Randomized relational decision trees.

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

the terminal distributions it encounters in the trees and taking the mode. In Amit et al. (1997) the application of this approach to the recognition of handwritten digits is described. We report a classification rate of 99.2% on the NIST database. In Amit & Geman (1997) the application of this approach to the recognition of hundreds of shape classes is analyzed, together with some theoretical aspects of the multiple randomized tree algorithm. This theoretical analysis is continued in Amit et al. (1999).

To summarize objects are classified in terms of a large pool of very sparse spatial arrangements of local image features. The pool is efficiently accessed using decision trees.

1.2 Extensions to speech recognition and theoretical analysis

The relational decision tree paradigm has been successfully applied to the recognition of isolated spoken digits. The features are very simple functions of the spectrogram, and the relations are temporal. Training on relatively small training sets this approach has achieved higher recognition rates than state of the art Hidden Markov Model methods. This is in the constrained situation of isolated utterances. The issue of combining recognition with segmentation has yet to be addressed in this context. However due to the use of multiple trees and the large degree of invariance incorporated in the relations the method is very robust to errors in segmentation. This is demonstrated in Amit & Murua (1999) by testing on randomly truncated versions of the data.

The use of multiple classifiers and in particular multiple decision trees has become a very powerful tool. See for example Breiman (1998), Breiman (1999), Schapire et al. (1998), Dietterich (1998). There are two complimentary methods for creating multiple classifiers. The first is randomization, for example the features employed for a split in the tree or the architecture of the network. The second is boosting where higher weights are given to data points which are misclassified by the current set of classifiers. In Amit et al. (1999) we attempt to provide a unifying explanation for the role of these two approaches as methods for *conditional covariance minimization* between the classifiers. Boosting and randomization are shown to both be sampling techniques from a distribution on the space of classifiers determined by the protocol and the training set. Certain simple moments of this distribution seem to determine the performance of the aggregate classifier.

1.3 Object detection

Spatial arrangements of local image features have also been used for an efficient object detection algorithm. In the previous proposal we suggested an approach to model registration based on decomposable graphs of local image features. The cliques of the graphs were triangles and a cost function was associated to each triangle penalizing its deviation from the model triangle on the template graph. All features are found in the image and a dynamic programming algorithm on the decomposable graph yields the optimal match. See Amit & Kong (1996) and Amit (1997) where the ideas are applied to automatic anatomy detection.

In comparison to elastic deformable template methods, this model represents a significant simplification of the underlying graph and of the associated computation. A sparse decomposable graph replaces the lattice type graph underlying the elastic models. The computation changes from continuum based gradient descent type algorithms to discrete dynamic programming. Due to these simplifications the output of the algorithm only consists of the match of a small number of model points, not of the entire object, thus we obtain less information on the instantiation of the object. The graphical models do not require initialization, and hence provide a crucial initial step for the elastic matching methods. Another useful attribute of the sparse graphical model is the possibility to explicitly impose constraints on certain deformations based on prior knowledge. However the model still needs to be constructed by hand and slows down considerably in complex scenes with a large degree of clutter or confusing background. In addition the model fails in the presence of occlusion.

This motivates yet a further simplification of the decomposable graph, to even simpler graphs, providing even simpler output: object location. In Amit (1998) the graph has been reduced to the simplest form possible while maintaining some form of constraints on the spatial arrangement of the local features: a star type graph. All features are constrained to lie in certain regions relative to a virtual center, see Figure 1.

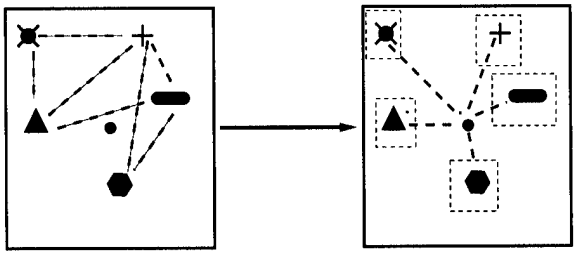


Fig. 1

The regions are set so that the virtual center of any instantiation consistent with the more complex graph, would be detected. It is important that the simpler models detect any instance the more complex models would detect, namely they should be *invariant* to instantiation parameters of the more complex models which are not part of the output of the simpler model.

Surprisingly the statistics of real images are such that using a star type graph with a moderate number of local features one obtains a very low number of false negatives at the price of only a few false positives. Moreover the graph is no longer required to be present in full as in the decomposable case. Rather any subset of sufficient size is sufficient to call a detection. This provides substantial robustness to occlusion. The training stage is fully automatic once the training images of the object are registered to a fixed scale and location on a reference grid.

A large pool of N local features consisting of flexible edge arrangements is predefined. A center edge and several other edges allowed to float in small regions around the center. On the right we show an example of a definition of a feature with three edges around the center. The feature is present at a location in the image if the center edge type is found there and each of the other edges is found anywhere in the respective region relative to the center.

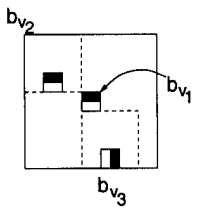


Fig. 2

In each small region of the reference grid a greedy search is carried out in this pool for a feature with high frequency (say greater than 50%) on the registered training images.

Typically several tens of locations are thus identified and a fixed size random subset of say $n = 20$ is chosen.

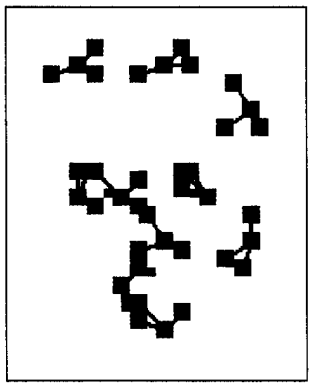


Fig. 3

The collection of identified features/location pairs is the object representation. The associated graph has an edge between each feature location z and the center of the reference grid. The edges are labeled by the region B_z in which the feature is allowed to float around the model location. This region is determined for example by the range of scales and rotations at which the object is to be detected, as well as other deformations allowed by the more complex models. See Figure 1 (right.) In Figure 3 are some of the features identified for faces at their location on the reference grid. The center edge of each feature defines its ideal location.

Given an image, all instances of each of the model features are identified. A search for virtual centers in the image where a sufficiently large number, say τ, of model features is present, within each respective region, is implemented using a generalized Hough transform. This is ideally suited for detecting the centers of such star type graphs. Each detected feature 'votes' for a region of centers consistent with the location of the feature in the model. In Figure 4 we show how any subgraph of size 3 of the star graph of Figure 1 is found.

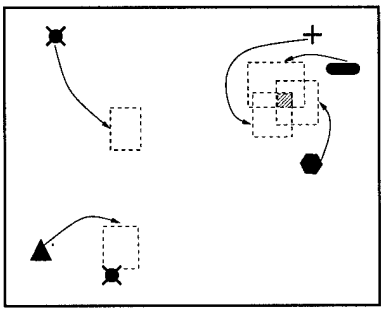


Fig. 4

Scale and rotation are subsequently identified at the candidate locations. Other more intensive graph matching computations, including elastic template matching can be carried out as well. These serve both to provide additional information on the instantiation of the object, as well as a filter on the false positives. If the match of the complex model has too high a cost, the detection is rejected. To find the object at significantly larger scales the image is reprocessed at several lower resolutions. This approach has been successfully implemented for face detection, symbol detection, and detection of 2d views of rigid objects. See Figure 5 for face detections, where scale and rotation are estimated and employed as a rejection mechanism. Computation time on a standard Pentium II is around 1.5 seconds for a 240x320 image processed as 6 resolutions. Figure 6 shows 'paper binder' detections with the same procedure.

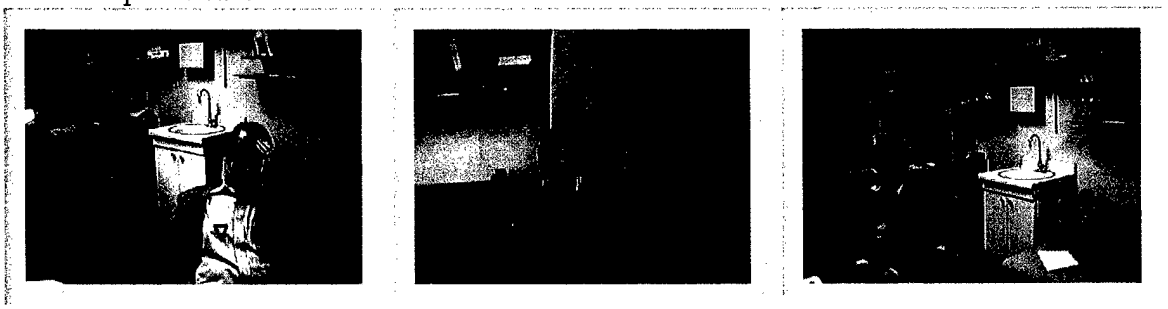


Fig 5.

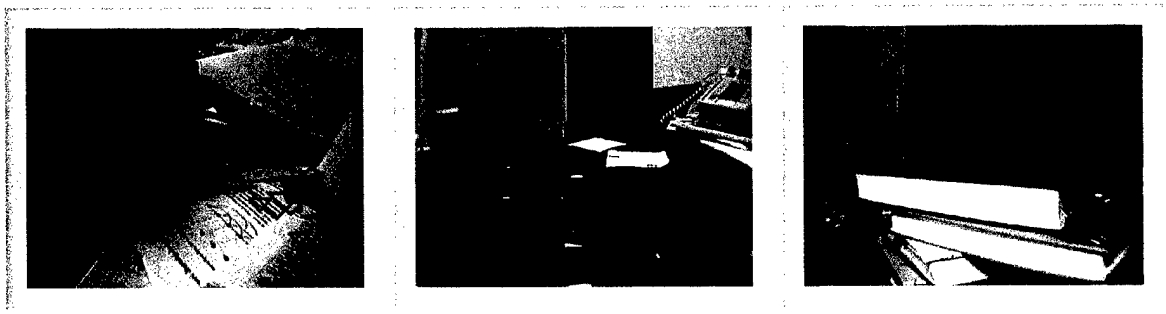


Fig 6.

The idea of learning more complex features and then represent an object in terms of a graph describing their geometric arrangement can be found in Burl et al. (1995), Wiskott et al. (1997), Cootes & Taylor (1996). In these approaches the features, or certain relevant parameters, are also identified through training. One clear difference however is that the approach presented here makes use only of binary features with hardwired invariances with well understood statistical properties, and employs a very simple form spatial arrangement for the object representation. This leads to efficient implementations of the detection algorithm.

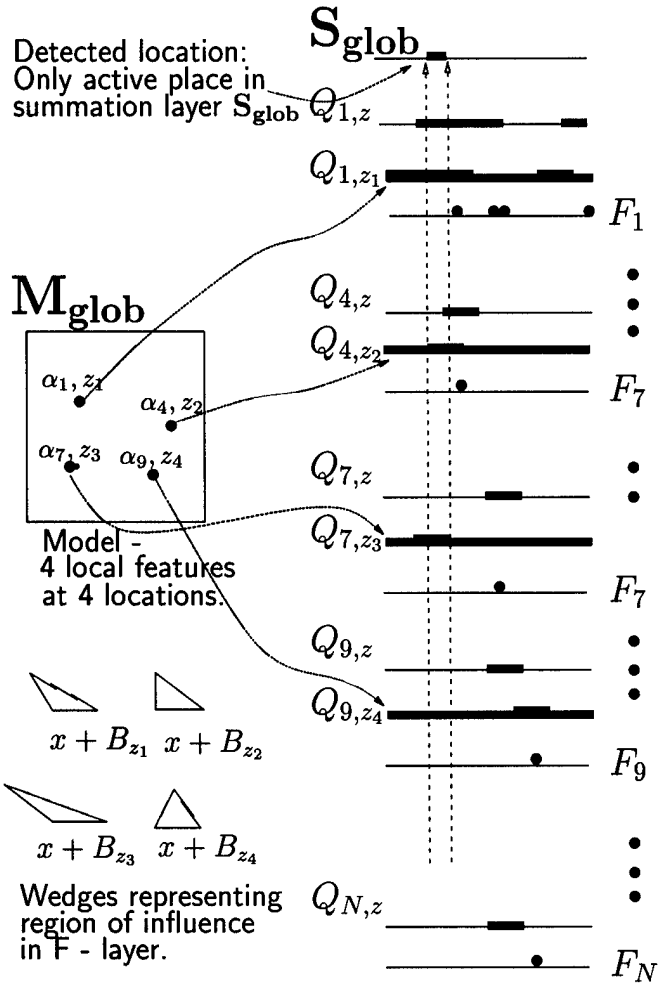
1.4 Neural network implementations

An attractive feature of the algorithm outlined above for training and detecting objects is the possibility of implementing it in a neural network with biologically plausible architecture, computation, and learning mechanisms. In Amit (1998) we propose a network which involves only binary neurons and is capable of detecting the object anywhere in the visual field by implementing the Hough transform for any object representation evoked in a central memory module. The network does not change its architecture for detecting new objects, it employs top-down priming in order to direct the bottom up flow of information (edge and local feature maps) in such a way that the Hough transform is computed in a simple retinotopic summation layer.

Define a module M_{glob} which has one unit corresponding to each feature/location pair. An object representation, consists of a small collection of n such pairs $(\alpha_{i_j}, y_j), j = 1, \ldots, n$. For simplicity assume n is the same for all objects, and that the number τ needed for a detection is the same for all objects as well. *An object representation is a simple binary pattern in M_{glob}, with n 1's and $(N - n)$ 0's.* Instances of each of the N local features is detected in an array F_i. For each i, introduce a system of arrays $Q_{i,z}$ indexed by the locations z in the reference grid G. These Q arrays lay the ground for the detection of *any* object representation by performing the 'voting' step of the Hough transform. Thus a unit at location $x \in Q_{i,z}$ receives input from a region $x + B_z$ in F_i.

Note that for each unit $u = (\alpha_i, z)$ in M_{glob} there is a corresponding $Q_{i,z}$ array. All units in $Q_{i,z}$ receive input from u. This is where the top-down flow of information is achieved: In order to be activated a unit x in $Q_{i,z}$ needs both (α_i, z) and some unit in the region $x + B_z$ in F_i to be activated. Thus the representation evoked in M_{glob} *primes* the appropriate $Q_{i,z}$ arrays to a point where they could be activated if the appropriate input comes from below,

i.e. the F_i array. The system of Q arrays sum into an array S_{glob}. A unit at location $x \in S_{glob}$ receives input from all $Q_{i,z}$ arrays at location x and is onif $\sum_{i=1}^{N} \sum_{z \in G} Q_{i,z}(x) \geq \tau$. The array S_{glob} therefore shows those locations for where there are more than τ votes in the Hough transform. In the figure below we provide a graphic representation of of this net. A more sophisticated network involving adaptable local feature detectors is also suggested in Amit (1998) as well some interesting biological analogies.



In this example the object is defined in terms of feature/location (α_{i_j}, z_j) pairs on the reference grid coded by a unit in M_{glob}. Three of the four have to be present to have a detection. Each unit provides input to all units in the corresponding $Q_{i,z}$ array (thick lines). The locations of the bottom-up feature detections are shown as •'s on the F_i arrays. They provide input to 'displaced' regions in the Q arrays shown as thick lines. The regions are define in terms of neighborhoods B_z of the model location z. Only locations on the Q array which receive input both from the F arrays and from a model unit in M_{glob} - double thick lines - is actually on and contributes to the summation into S_{glob}. Note that instances of feature α_9 do not contribute to the detection since they are is not present in the correct location relative to the others. There are N systems of F, Q arrays one for each local feature α.

To our knowledge there is no alternative network in the literature for translation invariant detection of objects which is not wired for a *specific* object representation. In Amit & Mascaro (1999) we describe a neural network architecture which employs local Hebbian learning both to achieve object recognition and to create object representations which can drive the detection network.

2 Publications

1. Amit, Y. & Kong, A. (1996), 'Graphical templates for model registration', *IEEE PAMI* **18**, 225–236.

2. Amit, Y., Geman, D. & Wilder, K. (1997), 'Joint induction of shape features and tree classifiers', *IEEE Trans. on Patt. Anal. and Mach. Intel.* **19**, 1300–1306.

3. Amit, Y. & Geman, D. (1997), 'Shape quantization and recognition with randomized trees', *Neural Computation* **9**, 1545–1588.

4. Yoshida, H., Katsuragawa, S., Amit, Y. & Doi, K. (1997),' Wavelet snake for classification of nodules and false positives in digital chest radiographs', *Proc IEEE Engineering in Medicine and Biology Society (IEEE-EMBS)*, 509–512.

5. Yoshida, H., Katsuragawa, S., Amit Y. & Doi K., (1997), 'Wavelet snake for classification of nodules and false positives in digital chest radiographs', *Proc. SPIE 3169: Wavelet Applications in Signal and Image Processing V*, 328–337.

   Amit, Y. (1997), 'Graphical shape templates for automatic anatomy detection: application to mri brain scans', *IEEE Trans. Medical Imaging* **16**, 28–40.

6. Amit Y. & Geman D. (1998), 'Discussion of 'Arcing Classifiers' by Leo Breiman', *The Annals of Statistics*, **26**.

7. Amit, Y. (1998), 'Deformable templates for object detection', *Tutorial Notes for the IEEE International Conference on Image Processing.*

8. Amit Y. and Geman D. (1999), 'A computational model for visual selection', *Neural Computation*, **11, 1691–1715.**

9. Amit, Y. & Mascaro, M. (1999), Modified hebbian learning on single layer attractor networks: an application to character recognition, Technical report, Department of Statistics, University of Chicago.

10. Amit, Y. & Murua, A. (1999), Speech recognition using randomized relational decision trees, Technical report, Department of Statistics, University of Chicago.

11. Amit, Y., Blanchard, G. & Wilder, K. (1999), Multiple randomized classifiers: MRCL, Technical report, University of Chicago.

12. Amit, Y. (2000), A neural network architecture for visual selection, *Neural Computation*, **To appear**.

3 Scientific Personnel

Graduate students

- Steve Wang - Ph. D. 1998.
- Gilles Blanchard - 1998-1999 (Visiting student from Ecole National Superieur).

Post-docs

- Bruno Jedynak. 1996-1997.
- Ken Wilder, 1997-1999.
- Massimo Mascaro, 1999-

References

Amit, Y. (1997), 'Graphical shape templates for automatic anatomy detection: application to mri brain scans', *IEEE Trans. Medical Imaging* **16**, 28–40.

Amit, Y. (1998), A neural network architecture for visual selection, Technical Report 474, Dept. of Statistics, University of Chicago.

Amit, Y. & Geman, D. (1997), 'Shape quantization and recognition with randomized trees', *Neural Computation* **9**, 1545–1588.

Amit, Y. & Kong, A. (1996), 'Graphical templates for model registration', *IEEE PAMI* **18**, 225–236.

Amit, Y. & Mascaro, M. (1999), Modified hebbian learning on single layer attractor networks: an application to character recognition, Technical report, Department of Statistics, University of Chicago.

Amit, Y. & Murua, A. (1999), Speech recognition using randomized relational decision trees, Technical report, Department of Statistics, University of Chicago.

Amit, Y., Blanchard, G. & Wilder, K. (1999), Multiple randomized classifiers: Mrcl, Technical report, Department of Statistics, University of Chicago.

Amit, Y., Geman, D. & Wilder, K. (1997), 'Joint induction of shape features and tree classifiers', *IEEE Trans. on Patt. Anal. and Mach. Intel.* **19**, 1300–1306.

Breiman, L. (1998), 'Arcing classifiers (with discussion)', *The Annals of Statistics* **26**, 801–849.

Breiman, L. (1999), Random forests, random features, Technical report, University of California, Berkeley.

Burl, M. C., Leung, T. K. & Perona, P. (1995), Face localization via shape statistics, *in* M. Bichsel, ed., 'Proc. Intl. Workshop on Automatic Face and Gesture Recognition', pp. 154–159.

Cootes, T. F. & Taylor, C. J. (1996), Locating faces using statistical feature detectors, *in* 'Proc., Second Intl. Workshop on Automatic Face and Gesture Recognition', IEEE Computer Society Press, pp. 204–210.

Dietterich, T. G. (1998), An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization, Submitted for publication to Machine Learning.

Schapire, R. E., Fruend, Y., Bartlett, P. & Lee, W. S. (1998), 'Boosting the margin: a new explanation for the effectiveness of voting methods', *The Annals of Statistics* **26**, 1651–1686.

Wiskott, L., Fellous, J.-M., Kruger, N. & von der Marlsburg, C. (1997), 'Face recognition by elastic bunch graph matching', *IEEE Trans. on Patt. Anal. and Mach. Intel.* **7**, 775–779.