REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-08-2015	Final Report	1-Jun-2011 - 31-May-2015

4. TITLE AND SUBTITLE
Final Report: GBS: Guidance By Semantics-Using High-Level Visual Inference to Improve Vision-based Mobile Robot Localization

5a. CONTRACT NUMBER
W911NF-11-1-0090

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Jason Corso

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
State University of New York (SUNY) at Bu
Sponsored Projects Services
402 Crofts Hall
Buffalo, NY 14260 -7016

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)
58260-CS-YIP.24

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
The overall objective in guidance by semantics is to improve the sensing and actuation of unmanned and optionally manned platforms by incorporating high-level visual inference into the robot loop. Our specific goal in this project is to perform mapping and localization on a mobile platform using semantically meaningful sensor data. In our case, we have used a camera image co-registered with a laser scan to filter in scan points that fall on buildings in the scene or an RGBD sensor. This enables the robotic platform to only make
use of readings that are known to be good, static landmarks, such as buildings as we have done in previous years

15. SUBJECT TERMS
mobile robotics, computer vision, visual inference, mobile robot guidance

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Venkat Krovi
UU	UU	UU	UU		19b. TELEPHONE NUMBER 716-645-1430

Report Title

Final Report: GBS: Guidance By Semantics-Using High-Level Visual Inference to Improve Vision-based Mobile Robot Localization

ABSTRACT

The overall objective in guidance by semantics is to improve the sensing and actuation of unmanned and optionally manned platforms by incorporating high-level visual inference into the robot loop. Our specific goal in this project is to perform mapping and localization on a mobile platform using semantically meaningful sensor data. In our case, we have used a camera image co-registered with a laser scan to filter in scan points that fall on buildings in the scene or an RGBD sensor. This enables the robotic platform to only make use of readings that are known to be good, static landmarks, such as buildings as we have done in previous years. However, buildings are just one small type of semantic inference we can and plan to use.

To realize these goals, in previous years, we have investigated full scene semantics with a rich parts based segmentation model, stairway detection, human motion and activity understanding, as well as more fundamental work on learning nonparametric distance functions to support more capable semantic inference across the different types of models. In the previous year, we have investigated full ascending stairway modeling, object-category sensing, multiple robot collaboration, and made strong progress toward the full guidance by semantics vision. The building and stairway detection routines have been implemented and deployed on ARL Packbots and ClearPath Turtlebots, with our ARL CISD colleagues and tested in the field at Camp Lejeune.

In the current year, we brought our semantic inference work up to a mature state and we studied efficient mapping using modern approximate graph inference techniques to enable us to conduct the final guidance by semantics experiments in the next and final year of the project.

We cover and describe all of this work in the report; more information can readily be provided if it is needed.

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
10/05/2012	1.00	Jason J. Corso. Toward Parts-Based Scene Understanding with Pixel-Support Parts-Sparse Pictorial Structures, Pattern Recognition Letters, (01 2013): 0. doi:
TOTAL:	**1**	

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
08/13/2013	13.00 Pradipto Das, Chenliang Xu, Richard F. Doell, Jason J. Corso. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching, IEEE Conference on Computer Vision and Pattern Recognition. 23-JUN-13, . : ,
08/13/2013	17.00 Jeffrey A. Delmerico, Jason J. Corso, David Baran, Philip David, Julian Ryde. Ascending Stairway Modeling: A First Step Toward Autonomous Multi-Floor Exploration, IEEE/RSJ Intelligent Robots and Systems (Video Proceedings). 07-OCT-12, . : ,
08/13/2013	15.00 Vikas Dhiman, Julian Ryde, Jason J. Corso. Mutual Localization: Two Camera Relative 6-DOF Pose Estimation from Reciprocal Fiducial Observation, IEEE/RSJ International Conference on Intelligent Robots and Systems . 05-NOV-13, . : ,
08/13/2013	14.00 Jeffrey A. Delmerico, David Baran, Philip David, Julian Ryde, Jason J. Corso. Ascending Stairway Modeling from Dense Depth Imagery for Traversability Analysis, IEEE International Conference on Robotics and Automation. 06-MAY-13, . : ,
08/25/2015	21.00 Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, Jason Corso. Can humans fly? Action understanding with multiple classes of actors, IEEE Conference on Computer Vision and Pattern Recognition. 08-JUN-15, . : ,
08/25/2015	23.00 Vikas Dhiman, Aghijit Kundu, Frank Dellaert, Jason Corso. Modern MAP inference methods for accurate and fast occupancy grid mapping on higher order factor graphs, International Conference on Robotics and Automation. 31-MAY-14, . : ,
08/25/2015	22.00 Jiasen Lu, Ran Xu, Jason Corso. Human Action Segmentation with Hierarchical Supervoxel Consistency, IEEE Conference on Computer Vision and Pattern Recognition. 08-JUN-15, . : ,
10/03/2014	18.00 Wei Chen, Caiming Xiong, Ran Xu, Jason J. Corso. Actionness Ranking with Lattice Conditional Ordinal Random Fields, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 23-JUN-14, Columbus, OH, USA. : ,
10/03/2014	19.00 Vikas Dhiman, Adhijit Kundu, Frank Dellaert, Jason Corso. Model MAP inference methods for accurate and faster occupancy grid mapping on higher order factor graphs, IEEE International Conference on Robotics and Automation. 31-MAY-14, . : ,
10/05/2012	7.00 Caiming Xiong, David M. Johnson, Ran Xu, Jason J. Corso. Random forests for metric learning with implicit pairwise position dependence, 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 12-AUG-12, Beijing, China. : ,
10/05/2012	8.00 Ran Xu, Priyanshu Agarwal, Suren Kumar, Venkat N. Krovi, Jason J. Corso. Combining Skeletal Pose with Local Motion for Human Activity Recognition, Proceedings of VII Conference on Articulated Motion and Deformable Objects. 28-JUL-12, . : ,
10/05/2012	9.00 Chenliang Xu, Jason J. Corso. Evaluation of super-voxel methods for early video processing, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 16-JUN-12, Providence, RI. : ,
10/05/2012	2.00 Julian Ryde, Jason J. Corso. Fast Voxel Maps with Counting Bloom Filters, IEEE/RSJ International Conference on Intelligent Robots and Systems. 08-OCT-12, . : ,

10/05/2012 3.00 Jason J. Corso, S. Sadanand. Action bank: A high-level representation of activity in video, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 16-JUN-12, Providence, RI. : ,

10/05/2012 4.00 Caiming Xiong, Jason J. Corso. Coaction discovery: segmentation of common actions across multiple videos, Proceedings of the Twelfth International Workshop on Multimedia Data Mining (MDMKDD). 12-AUG-12, Beijing, China. : ,

10/05/2012 5.00 Caiming Xiong, David M. Johnson, Jason J. Corso. Spectral Active Clustering via Purification of the k-Nearest Neighbor Graph, European Conference on Data Mining. 14-JUL-11, . : ,

10/05/2012 6.00 Caiming Xiong, David M. Johnson, Jason J. Corso. Efficient Max-Margin Metric Learning, European Conference on Data Mining. 14-JUL-11, . : ,

10/05/2012 10.00 Chenliang Xu, Caiming Xiong, Jason J. Corso. Streaming Hierarchical Video Segmentation, Proceedings of European Conference on Computer Vision. 08-OCT-12, . : ,

10/05/2012 11.00 Jeffrey. A. Delmerico, Philip David, Jason. J. Corso. Building facade detection, segmentation, and parameter estimation for mobile robot localization and guidance, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011). 25-SEP-11, San Francisco, CA. : ,

TOTAL: **19**

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

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received	Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Jason Corso	0.08	
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
Jeff Delmerico
Total Number: 1

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

This list is continuous from the project inception.

Semantic Filtering. This system has been implemented and was deployed on a Packbot during Jeffrey Delmericos summer internship at ARL in 2011, and tested more extensively during field tests with ARL at Camp Lejeune in October 2011. More data acquisition was performed at Camp Lejeune in September 2012.

Stairway Detection. This system has been implemented and was deployed on a Packbot during Jeffrey Delmericos summer internship at ARL in 2011, and tested more extensively during field tests with ARL at Camp Lejeune in October 2011 and September 2012. The detector exhibited very high accuracy and a negligible false-positive rate. The assessment routines also demonstrated quantitatively high accuracy for traversability (e.g., within a few degrees error incline).

Action Bank. The human activity recognition system developed in 2011 year, Action Bank, has received considerable commercial interest. The university has filed a patent in Dec. 2012 on it.

Jason J. Corso
SUNY Buffalo / University of Michigan
Grant No.: W911NF-11-1-0090

Project Years: 2011–2015
PI: V. Krovi / J. Corso
Prepared on: August 25, 2015

FINAL PROGRESS REPORT ATTACHMENT
Guidance By Semantics—Using High-Level Visual Inference to Improve Vision-based Mobile Robot Localization
SUNY Buffalo / University of Michigan

This document contains material for the cumulative project period. The material presented is fully or partially supported by this grant.

This project was initially awarded to SUNY Buffalo with J. Corso as PI while he was on the faculty at SUNY Buffalo. He subsequently moved to the University of Michigan. SUNY at Buffalo appointed V. Krovi as PI and the entire remainder of the award was subcontracted out to the University of Michigan with J. Corso as PI.

1 Objective

Unmanned and optionally manned mobile ground systems, such as the Packbot, have a critical need to autonomously localize themselves in diverse environments, including indoor and outdoor environments in which external positional sensors, like GPS, fail to provide consistent, spatially resolved localization, especially in "urban canyons." Vision-based methods are the main solution, but existing methods rely primarily on bottom-up features that may be found in undesirable, non-stationary, reflective, etc. positions in the scene. The project emphasizes the use of high-level scene and object semantics, which are automatically inferred from the visual data using state of the art computer vision and machine learning ideas, to guide the selection of these features and other parameters of the vision-based UMS methods. Ultimately, this work is expected to improve the speed, accuracy, and robustness of the robotic ground systems to help our soldiers better achieve their missions. There is rich potential in improving the accuracy and robustness of vision-based localization and visual odometry with broad Army and DoD relevance from our Guidance By Semantics—GBS—methodology.

2 Activities and Findings

2.1 Overall Approach in Semantic Filtering

This section describes the overall approach in the project and leads into the specific research activities we have conducted.

The overall objective in guidance by semantics is to improve the sensing and actuation of unmanned and optionally manned platforms by incorporating high-level visual inference into the robot loop. Our specific goal in this project is to perform mapping and localization on a mobile platform using semantically meaningful sensor data. In our case, we have used a camera image co-registered with a laser scan to filter in scan

points that fall on buildings in the scene or an RGBD sensor. This enables the robotic platform to only make use of readings that are known to be good, static landmarks, such as buildings as we have done in previous years. However, buildings are just one small type of semantic inference we can and plan to use.

To realize these goals, in previous years, we have investigated full scene semantics with a rich parts-based segmentation model, stairway detection, human motion and activity understanding, as well as more fundamental work on learning nonparametric distance functions to support more capable semantic inference across the different types of models. In the previous year, we have investigated full ascending stairway modeling, object-category sensing, multiple robot collaboration, and made strong progress toward the full guidance by semantics vision. The building and stairway detection routines have been implemented and deployed on ARL Packbots and ClearPath Turtlebots, with our ARL CISD colleagues and tested in the field at Camp Lejeune.

In the current year, we brought our semantic inference work up to a mature state and we studied efficient mapping using modern approximate graph inference techniques to enable us to conduct the final guidance by semantics experiments in the next and final year of the project.

We cover and describe all of this work in the report; more information can readily be provided if it is needed.

2.2 Major Research and Education Activities

This has been an active year for our work on the grant. We summarize the major research activities in the following list, which is in reverse chronological order.

Guidance by Semantics. We have conducted a full range of experiments toward understanding how visual semantic filtering impact the quality of vision-based robot localization.

Two-robot mutual localization and 6-DOF pose estimation. We have extended the space of robots in study by moving from one robot navigation to pairs of robots jointly navigating. Currently, we have developed a generalization of the Perspective-3-Points problem where the observer and the observed points are distributed across different reference frames.

Actionness Ranking carefully analyzes the notion of action as a distinct type of motion signal. This work extends pushes the action recognition community beyond direct classification of data-specified human actions into one of parsing a video into regions of likely action, which can then later be classified into specific actions. We propose a new ordinal random field framework to approximately infer an ordering of video regions according to the actionness.

Fast, Approximate Inference in Mapping explores how the forward sensor model can be used in occupancy grid mapping without expensive optimization routines like Expectation Maximization and Gibbs sampling. Using the inverse sensor model has been popular in occupancy grid mapping. However, it is widely known that applying the inverse sensor model to mapping requires certain assumptions that are not necessarily true. Even the works that use forward sensor models have relied on methods like expectation maximization or Gibbs sampling which have been succeeded by more effective methods of maximum a posteriori (MAP) inference over graphical models. In this work, we propose the use of modern MAP inference methods along with the forward sensor model. Our implementation and experimental results demonstrate that these modern inference methods deliver more accurate maps, more efficiently than previously used methods.

Ascending stairway modeling and traversability assessment. We have developed a stairway detecting and modeling system that runs in real-time and has demonstrated robust performance on Army-relevant sites. The goal is to use geometric cues from depth data to detect ascending stairwells and operate at the frame rate of the sensor when deployed on a mobile platform. Detection of stairwells by mobile robots will enable multi-floor exploration and mapping for those platforms capable of stair traversal. We demonstrate results on this in detail below.

Attributed object maps with object sensing while mapping. We have extended the range of classes we use for semantic filtering to include indoor small objects, such as boxes. We have developed a system that uses RGBD sensing and joint recognition and reconstruction of such object categories.

Semantic visual inference for mobile robot navigation. We have been exploring the role that the visual inference methods in study in this project can play in the application domain of mobile robot simultaneous mapping and navigation. We have heavily focused on building facades and ascending stairways during this past year. In the future, we plan to extend these to comprehensive visual semantic mapping for mobile robots.

Mid-level representations for Moving Platforms—Video. Video understanding research has focused on low-level features for representation: points and trajectories. Although substantial progress has been born of these features, it is non-trivial to either transform these point-based mechanisms into spatiotemporal segmentations or infer much about the high-level semantics they are contain. Like superpixels have gained strong support in the image understanding community, we contend that supervoxels have rich potential to provide an alternative representational fabric for video understanding, one that is similarly rapidly computed, has a natural graph-structure, can represent spatiotemporal regions, and has potential to better transfer semantics to subsequent processes. Our two papers on this topic have laid the groundwork for this effort [20][21]. The code is openly and freely available.

Semantics of humans moving in the scene. Video-based activity recognition is dominated by a low-level feature extraction mechanism followed by a standard pattern recognition classifier; these methods work well in the face of the high-dimensional and complex video-activity signal. However, it is hard to transfer the semantic from these low-level classifiers to subsequent inferential processes; it is also hard (or not possible) to precisely understand why a particular feature method is unable to accurately recognize a particular action. We, in contrast, have proposed a representation for activity recognition based on the responses of several hundred action templates, wherein the action semantics directly comprise the *feature* on which the activity recognition mechanisms are based [14]. Our performance dominates the low-level feature-based approaches on all major benchmarks attempted. Our code is available and released to the community.

Metric Learning for semi-supervised clustering in heterogeneous sample spaces. Images and videos live in an inconceivably high-dimensional sample space. The majority of successful methods to date for elements of image and video understanding within the computer vision community are based on supervised learning, which, although has demonstrated success in select problems, does not scale to the broader problem of visual understanding. We have been exploring the role the metric learning can play to provide a mid-level representation on which higher-level concepts can be learned without full annotation required. We proposed two distinct metric learning methods for these purposes [17] [18], one of which won a Best Paper award [17].

2.3 Major Findings

Each major finding is demarcated by a horizontal rule, and a large, boldface heading.

Guidance by Semantics

Early Work with ARL on Data Capture for GBS We have made strong progress toward the overall goal of guidance by semantics. This work emphasizes semantic filtering on a point cloud and heavily utilizes our other developments in semantic video understanding. The basic intuition is to filter out classes of objects/scenery that are undesirable, and filter in those classes that will be useful for some subsequent task. The specific problem we consider is 3D mapping in a cluttered and potentially dynamic environment. In this case, we seek to extract the points from a series of point clouds that correspond to buildings and structures, while removing other objects. Buildings are essentially static objects, so they make for reliable landmarks for mapping, whereas other objects in the scene may not have the same level of permanence or may not be good landmarks. For example, a car that is parked in front of a building may park across the street in the future. Our goal is to use point clouds that are registered with a camera to perform visual classification of each point, and then filter the point cloud based on this classification. We use a robot setup with a panoramic camera and a 2D laser scanner on a 1-axis nodder, producing 3D point clouds when assembled over a period of time.

With our ARL collaborators, we collected a large volume of data at Camp Lejeune in September 2012 using such a setup. We are currently considering only binary classification into classes of building and background. We then intend to show that it is possible to construct a map using just the static elements of the scene with the dynamic elements removed, and hopefully that this map may be more accurate because it uses only reliable landmarks. This is a work in progress and we are currently developing a classifier that provides labels for the scene, and effectively isolates the points on buildings. We are now able to accurately filter in buildings, and filter out the ground, but more work needs to be done in segmenting and filtering out clutter. In particular, there is considerable bias in the training data, such that the background class is represented significantly more with the ground and the sky than any clutter. Therefore, the model learned for the background is only effective at labeling the ground and sky, and not the clutter. Relabeling the training data in order to balance the representation of the clutter in the background model may resolve this issue.

GBS Experiment Hypothesis Visual odometry is based on the assumption of static world and we hope that one can use semantics to detect common moving objects in a static world and improve performance by removing outliers lying on detected objects.

Experimental setup We used state of art implementation of visual odometry by Geiger et al. [GLSU13, GZS] released as a library called LIBVISO2. The library works by tracking blob and corner features across frames, followed by RANSAC based outlier removal while solving for ego-motion in a Gauss-Newton optimization process.

We used pedestrian and car datasets from KITTI [GLSU13] and ETH Mobile platform [BRL$^+$11] to test our hypothesis. We use the evaluation benchmark provided with the KITTI dataset. KITTI dataset is designed as a test bed for autonomous driving with ground truth odometry captured from GPS measurements.

ETH Mobile platform [BRL+11] dataset is multi-person tracking dataset with moving camera. The dataset includes wheeled odometry as ground truth data.

For semantics based outlier removal, we VOC-DPM library by Girshick et al. `http://people.cs.uchicago.edu/~rbg/latent-release5/` [FGMR10] to detect pedestrians and cars which cover almost all the non-stationary objects in our datasets. Once we detect the objects in a frame we reject all the tracked feature points lying within the bounding box. We compare the modified (with semantics) and unmodified versions (no semantics) of the LIBVISO2 algorithm on the chosen datasets.

Discussion The results (Figure 1) show that using object detections for outlier removal either deteriorates the performance or perturbs it by small amount. This can be explained due to robustness of RANSAC based outlier removal method. Since most of the effective outliers are already removed, hence object detections do not help. Moreover spurious detections end up removing inliers hence causing deterioration in performance. RANSAC based methods are expected to fail in a situation when an object or multiple objects, moving in the same direction,occupy majority of the visual field. This situation is very unlikely in natural scenes.

Our hypothesis of using semantics by employing object detections for outlier removal in visual odometry is not validated by experiments, especially when a reliable RANSAC based method is being used for outlier removal.

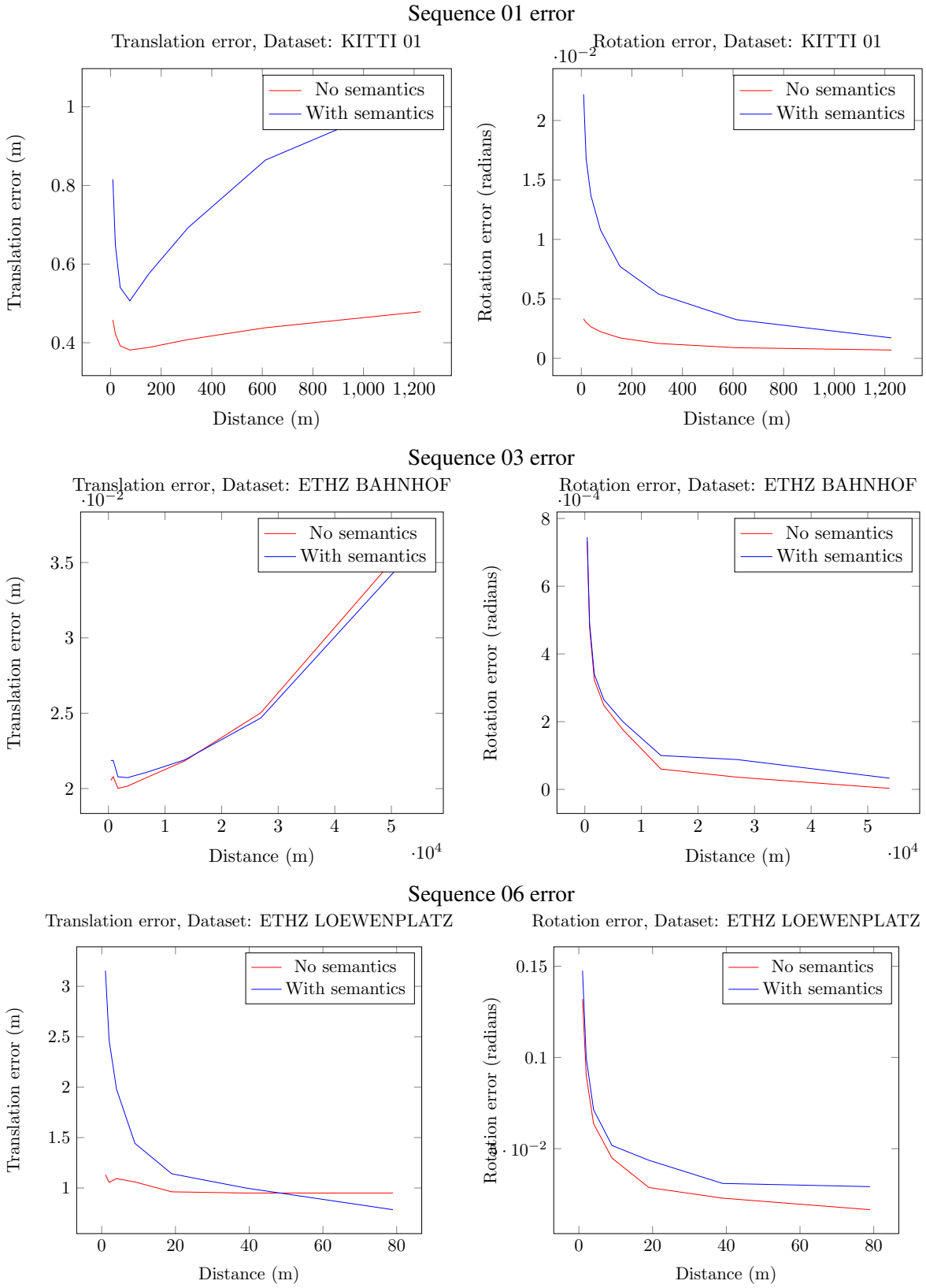

Figure 1: Experimental Results from our Guidance By Semantics Study.

Two-robot mutual localization and 6-DOF pose estimation.

Continued work in Cooperative localization occurred in the current year. Cooperative localization is the problem of finding the relative 6-DOF pose between robots using sensors from more than one robot. Various strategies involving different sensors have been used to solve this problem. For example, Cognetti et al. [CSF+12, FOS09] use multiple bearning-only observations with a motion detector to solve for cooperative localization among multiple anonymous robots. Trawny et al. [TZZR10] and lately Zhou et al. [ZR10, ZR12] provide a comprehensive mathematical analysis of solving cooperative localization for different cases of sensor data availability.

To the best of our knowledge, all other cooperative localization works require estimation of egomotion. However, a dependency on egomotion is a limitation for systems that do not have gyroscopes or accelerometers, which can provide displacement between two successive observations. Visual egomotion, like MonoSLAM [DRMS07], using distinctive image features estimates requires high quality correspondences, which remains a challenge in machine vision, especially in cases of non-textured environments. Moreover, visual egomotion techniques are only correct up to a scale factor. Contemporary cooperative localization methods that use egomotion [ZR10, TZZR10, Mar12] yield best results only with motion perpendicular to the direction of mutual observation and fails to produce results when either observer undergoes pure rotation or motion in the direction of observation. Consequently, in simple robots like Turtlebot, this technique produces poor results because of absence of sideways motion that require omni-directional wheels.

To obviate the need for egomotion, we have proposed and thoroughly experimentally validated a method for relative pose estimation that leverages distance between fiducial markers mounted on robots for resolving scale ambiguity. Our method, which we call *mutual localization*, depends upon the simultaneous mutual/reciprocal observation of bearing-only sensors. Each sensor is outfitted with fiducial markers (Fig. 2) whose position within the host sensor coordinate system is known, in contrast to assumptions in earlier works that multiple world landmarks would be concurrently observable by each sensor [ZT12]. Since our method does not depend on egomotion, hence it is instantaneous, which means it is robust to false negatives and is not susceptible to the errors in egomotion estimation.

The main contribution of our work is a generalization of *Perspective-3-Points* (P3P) problem where the observer and the observed points are distributed in different reference frames unlike conventional approaches where the observer's reference frame does not contain any observed points and vice versa. We have presented an algebraic derivation to solve for the relative camera pose (rotation and translation) of the two bearing-only sensors in the case that each can observe two known fiducial points in the other sensor; essentially giving an algebraic system to compute the relative pose from four correspondences (only three are required in our algorithm but we show how the fourth correspondence can be used to generate a set of

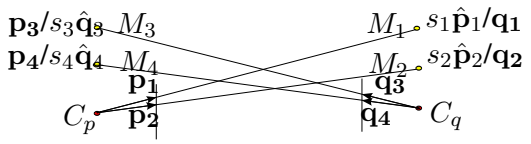

Figure 2: Simplified diagram for the two-camera problem. Assuming the length of respective rays to be s_1, s_2, s_3, s_4 respectively, each marker coordinates can be written in both coordinate frames $\{p\}$ and $\{q\}$. For example M_1 is $s_1\hat{\mathbf{p}}_1$ in frame $\{p\}$ and $\mathbf{q_1}$ in $\{q\}$, where $\hat{\mathbf{p}}_1$ unit vector parallel to $\mathbf{p}_1$.

hypothesis solutions from which best solution can be chosen). The details of the formulation are beyond the scope of this report but available in the reference IROS 2013 paper, below. Two fiducial points on each robot (providing four correspondences) are preferable to one on one and two on the other, as it allows extension to multi-robot (> 2) systems ensuring that any pair of similarly equipped robots can estimate their relative pose. We have focused on only the two robot case as an extension to the multi-robot case as pairwise localization is straightforward yet practically effective.

Our derivation, although inspired by the linear pose estimation method of Quan and Lan [QL99], is novel since all relevant past works we know on the P3P problem [HLON94], assume all observations are made in one coordinate frame and observed points in the other. In contrast, our method makes no such assumption and concurrently solves the pose estimation problem for landmarks sensed in camera-specific coordinate frames. We have implemented and tested our method on our mobile robot fleet and find it able to effectively perform 6-DOF mutual localization and hence lead to strong 3D reconstruction and mapping. Our paper reference below depicts these quantitative results.

This work was presented in November 2013 at IROS in Tokyo, Japan.

- V. Dhiman, J. Ryde, and **J. J. Corso**. Mutual localization: Two camera relative 6-dof pose estimation from reciprocal fiducial observation. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2013.

Actionness Ranking with Lattice Conditional Ordinal Random Fields

- W. Chen, C. Xiong, R. Xu, and **J. J. Corso**. Actionness ranking with lattice conditional ordinal random fields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

The computer vision community has achieved marked success in automatic action recognition (classification) from video (clips). Notable examples include the introduction of local action features with bags-of-words framework [WUK+09], such as spatio-temporal interest points [Lap05], trajectory-based representations [MPK09, WKSL11], and motion interchange patterns [KGGHW12] and the more holistic action bank representation which embeds a video into an action space by responses of individual action detectors [S**Corso**12]. These methods are enabling futuristic vision applications like automatic video-to-text [DXD**Corso**13, KMM+13] and smart classrooms [RX02].

However, in all of this so-called *action recognition* work in our field, the very notion of action has not been carefully defined or explicitly studied, although a hierarchy of actions and activities has been discussed [MHK06]. Instead, action is defined implicitly by examples in a dataset. UCF Sports [RAS08], for example, emphasizes olympic sports as action whereas HMDB51 [KJG+11] focuses more on everyday human actions such as brushing hair and hugging.

We carefully study the notion of action, leveraging ideas from the philosophy of action. There are four aspects to define *action* in the philosophy of action [Dav01]: first, action is what an **agent** can do; second, action requires an **intention**; third, action requires a **bodily movement** guided by an agent or agents; and fourth, action leads to **side-effects**. For example, playing with a ball is an instance of action. A person is able to play with a ball. Doing this action needs the movement of the human body; the person moves the ball by moving his or her hands and/or feet. When a person plays with a ball, a ball movement from left to right and up to down is just a side-effect since the ball has no intention. Its movement is barely the result of the action (playing) of the person. These phenomena are illustrated in Figure 3.

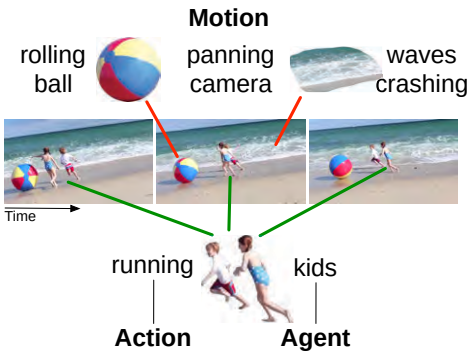

Figure 3: Our work distinguishes intentional action of an unknown agent (the kids in this example) from various other motions, such as the rolling ball, the crashing waves and the background motion from the panning camera.

Above, we highlighted the key words for the four aspects of action: agent, intention, bodily movement, side-effects. Two of these are directly observable in video: agent and bodily movement (in an image, one can only observe agent pose but not the bodily movement). Intention is not directly observable but not irrelevant from a computer vision point of view: a non-biological agent, such as a bicycle can not have intention, and hence the agents we care about are people and animals. Finally, side-effects may be directly observed in images, but these would involve a complex inference even farther beyond the reliable capability of our field than person and animal detection. Therefore, we define **actionness** as intentional bodily movement of biological agents. Actionness is a subclass of general motion and a direct presentation of action.

However, the capability of current person detectors and trackers are not sufficient to be fully leveraged as a basic for actionness extraction (our full analysis provides detailed quantitative results to this point). Therefore, our work seeks to extract a rank ordering of video regions according to the degree to which they

contain an action. We call this notion *actionness*. And, it is learned from data in a novel ordinal random field framework.

We target a rank ordering of actionness by regions rather than a direct classification of whether or not a region contains an action for two primary reasons. First, the foundational notion of action as an agent's intentional motion immediately presents a difficulty: agent (e.g, person, animal) detection remains a challenging and open problem [EVGW$^+$10]. There exist comparatively strong methods like deformable parts models [FGMR10], but the average precision remains too low for robust use (e.g., about 49.5 for person is the state of the art [GFM11]). Ranking makes it plausible to forego agent detection and segmentation prior to actionness classification; rather, directly ranking various regions of the image/video is more robust. Second, in any given image or video there may be more than one agent performing an action. Ranking which is more likely an action is hence more informative than simple classification.

We accomplish this ranking with a new model called the lattice conditional ordinal random field, which solves the linear ordering problem approximately using local features to score a given region and enforcing local ordering on the lattice.

We propose a conditional random field model M that captures the ordering as its random variables:

$$M_d(\{o_i\}_{i=1}^n|\mathcal{V},\mathcal{R},\theta) = \frac{1}{Z[\mathcal{R}]}\exp\left[\sum_i \alpha f_d(o_i,r_i) + \sum_{i,j}\beta g_d(o_i,o_j,r_i,r_j)\right], \tag{1}$$

where $Z[\mathcal{R}]$ is the normalization function and $\theta = (\alpha,\beta)$ are model parameters. Each o_i is the ordinal index of (given) region r_i; these indices take values from $\{1,2,\ldots,n\}$ and satisfy a strict ordering $o_1 > o_2 > \cdots > o_n$. Functions f_d and g_d capture the unary ordinal preference and pairwise ordinal agreement, which we will make explicit below.

Satisfying the strict ordering constraint on $\{o_i\}_{i=1}^n$ and the discrete nature of this ordering make learning and inference intractable. So, we relax our model to be a continuous CRF and replace o_i with a real-valued variable a_i for each region r_i. Furthermore, we relax the strict ordering to be a partial ordering such that $a_1 \geq a_2 \geq \cdots \geq a_n$. The relaxed model is written

$$M(\{a_i\}_{i=1}^n|\mathcal{V},\mathcal{R},\theta) = \frac{1}{Z[\mathcal{R}]}\exp\left[\sum_i \alpha f(a_i,r_i) + \sum_{i,j}\beta g(a_i,a_j,r_i,r_j)\right]\ . \tag{2}$$

Functions f and g are the continuous versions of f_d and g_d.

We adopt a generalized Hough voting framework for the unary ranking evidence term. The pairwise term uses an AdaBoost classifier to predict the local ranking preference of two regions and penalizes the current ranking when it violates the classifier prediction. To provide an effective situation for learning and inference, we relax the discrete ordering problem in the random field to a continuous one and derive exact solutions for inference and a gradient descent method for learning.

We have tested this model for actionness ranking on both image and video datasets and find it outperforms all baseline methods, including the recent Ranking SVM, which cannot enforce local ordering agreement.

Fast, Approximate Inference for Occupancy Grid Mapping

Mobile robot problems like navigation, path planning, localization and collision avoidance require an estimate of the robot's spatial environment; this underlying problem is called robot mapping[Thr02]. Even in environments in which maps are available, the environment may change over time necessitating a mapping ability on the mobile robot. Robot mapping hence remains an active field of research [MDBB12, NUS12, MB13] as it is an important problem in application areas like indoor autonomous navigation, grasping, reconstruction and augmented reality.

Although robot mapping can be performed in many ways—metric or topological; with range sensors, like sonar [Thr03], laser scanners [Thr03] and RGBD [NDI$^+$11], or bearing-only sensors [DRMS07, KKJ11]—metric mapping with range sensors is the most common. Bearing-only sensors provide estimates up to scale; topological maps still require local metric estimates for certain problems like navigation. We hence focus on metric mapping with range sensors, specifically, laser scanners.

Occupancy grid mapping (OGM) is a popular and useful range-based mapping method [Elf89, Mor88]. It affords a simple implementation and avoids a need to explicitly seek and match landmarks in the environment [Sug88, BG97]. In contrast, it discretizes the environment into cells, squares (2D) or cubes (3D), and associates a random variable with each cell that represents the probability of the cell being occupied or free.

OGM methods vary in how cell occupancy is estimated, but most methods make use of an inverse sensor model by assuming that the occupancy of each cell can be estimated independently of the other cells in the map [Elf89, Mor88, ME85, NDI$^+$11]. The main reason for using this independence assumption is computational efficiency. However, the assumption is inaccurate and can lead to overconfident estimates of occupancy in noisy observations [Thr03, MB13].

To overcome this limitation, Thrun [Thr03] proposes use of forward sensor model and expectation maximization to estimate occupancy. Following this line of work, more recently, Merali et al. [MB13] defines a Gibbs sampling algorithm based on a conditional estimate of cell occupancy given the rest of the map. Although these methods have relaxed the assumptions of independence, they remain computationally expensive and hence limited in applicability. For example, it is widely known that Gibbs sampling is computationally expensive and can get caught in local maxima [Liu02].

In contrast, in this work, we explore the use of modern inference algorithms for more effective occupancy grid mapping with forward sensor models. Our contributions are two fold thus far. Firstly, we introduce the factor graph approach to occupancy grid mapping problem, which, to the best of our knowledge, has not been applied to this problem. This factor graph formalism makes it plausible to apply modern fast inference algorithms, such as loopy belief propagation [KFL01] and dual decomposition [SGJ11].

Secondly, we introduce a class of higher order factors for our factor graph approach. Factor graph inference is exponential in neighborhood size, which requires us to focus on a certain sub-class of factors for tractability, such as the linear constraint-nodes [Pot07] or pattern-based factors [KP09]. We extend the pattern-based factors, which explicitly computes the potential only for certain factors matching a given set of patterns and otherwise assigns a constant. Whereas the pattern-based factors in [KP09] defines each pattern with a fixed value for each node, we generalize these pattern-based factors by allowing for *free* nodes whose value does not impact the computed marginal.

We have implement these contributions for effective occupancy grid mapping with a forward sensor model and test our work on both simulated and real-data. Our experiments demonstrate the effectiveness of our

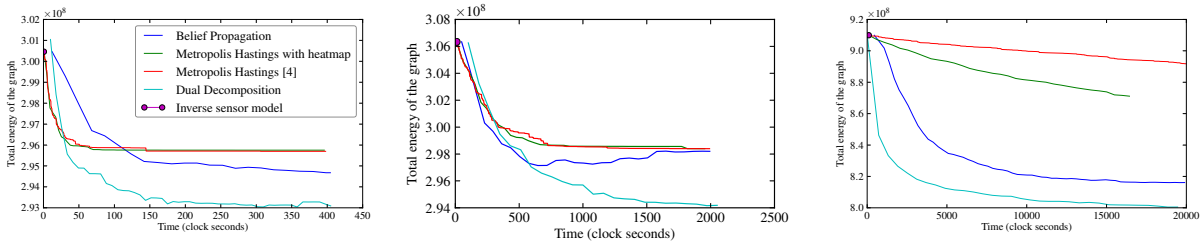

Figure 4: Comparison of convergence rate of different algorithms on occupancy grid mapping. The left graph shows the convergence on simulated dataset *cave*, while the right shows for *albert-b* [HR03] dataset. While sampling methods like Metropolis hastings converge quickly they stay far from optimum energy. On the other hand modern inference algorithms like belief propagation and dual decomposition reach closer to an optimum value. The legends are listed only once for clarity.

novel OGM approach, especially, dual decomposition. Details of our method can be found in the paper referenced below.

We discuss experiments on simulated as well real data. The simulated data is generated using *Player/Stage* [GVH03] project. We use multiple map bitmaps bundled along with player/stage library. The robot motion is generated using the wander driver. The robot is allowed to wander in the map for 2 minutes aggregating approximately 270,000 laser measurements.

For real data, we have used the *albert-b-laser* dataset provided by Cyrill Stachniss from University of Freiburg. The dataset was captured by a B21r robot with a SICK PLS moving through a lab at University of Freiburg. This data set was obtained from the Robotics Data Set Repository (Radish) [HR03]. We thank Cyrill Stachniss for providing this data.

To evaluate the convergence rate of each algorithm, we plot total energy (negative log likelihood) of the graph with respect to CPU ticks used by the algorithm. The plots of energy convergence with respect to time for cave dataset is shown in Fig. 4. This data clearly show the improvement from moving to belief propagation and dual decomposition, which, in all cases, leads to lower energies faster than our baselines. DD outperforms BP in typical cases.

In all our experiments we do not use any occupancy prior, although Merali et al. [MB13] suggest using an occupancy prior of 0.3 for better convergence. We use a step size of 50 for dual decomposition and piecewise constant sensor model. Also, we prefer piecewise constant sensor model over Gaussian sensor model because of the former being faster which is a consequence of having fewer patterns in the pattern based factor formulation. We have implemented the algorithms in C++ and the code will be released upon the acceptance of the paper.

As is evident from convergence comparison in Fig. 4, sampling algorithms (Metropolis Hastings with/without heatmap) are liable to getting stuck in a local minima. This is also an artifact of the simple transition probability where we flip only one cell at a time. Even from the qualitative results for sampling algorithms, we see that the walls are thinner than corresponding results in other algorithms which shows the inability of sampling based algorithms to form lower energy thicker walls for piecewise constant sensor model. The downside of being biased towards thinner walls is evident in the *albert-b* dataset (see Fig 5), as we get ragged walls for sampling algorithms.

- V. Dhiman, A. Kundu, F. Dellaert, and **J. J. Corso**. Modern MAP inference methods for accurate and

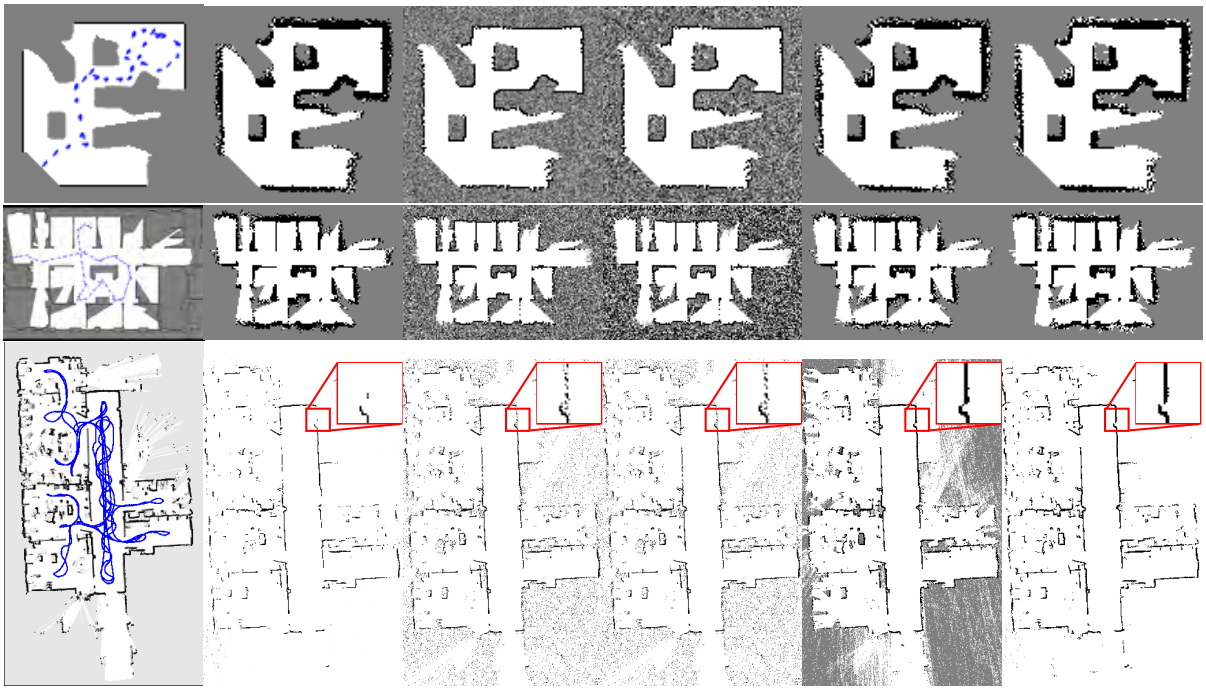

Figure 5: Qualitative results on different datasets. Each row represents a different dataset while each column represents a different algorithm. Th.e columns correspond to the following algorithms (from left to right): 1) Ground truth with the trajectory of the robot 2) Inverse sensor model 3) Metropolis Hastings without heat map 4) Metropolis Hastings with heat map 5) Belief Propagation (BP) 6) Dual decomposition (DD). The rows correspond to the following datasets (from top to bottom): 1) cave 2) albert-b. The grainy-ness in columns (3) and (4) is an artifact of sampling algorithms, when we sample over finite number of sample to compute expected state of a cell. Also note the missing or ragged walls in first 3 algorithms, while BP and DD are able to converge to thick solid walls. [HR03].

faster occupancy grid mapping on higher order factor graphs. In *Proceedings of International Conference on Robotics and Automation*, 2014.

Ascending Stairway Modeling and Traversability Estimation

The goal is to use geometric cues from depth data to detect ascending stairwells and operate at the frame rate of the sensor when deployed on a mobile platform. Detection of stairwells by mobile robots will enable multi-floor exploration and mapping for those platforms capable of stair traversal. Furthermore, stairways are fixed environment elements and can hence be used for semantic filtering.

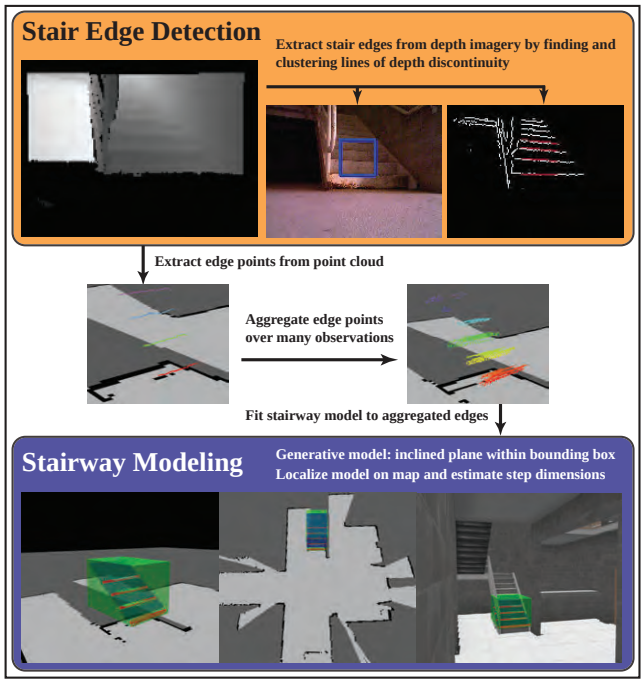

Figure 6: Workflow of the stairway detection and modeling system. Stair edges are extracted from depth imagery and aggregated over many observations. Periodically, a generative model of a stairway is fit to the aggregate cloud and its parameters re-estimated. The result is a model localized with respect to the robot's map of its environment.

Autonomous mobile robots have traditionally been restricted to exploring single floors of a building or outdoor areas that are free of abrupt elevation changes such as curbs and stairs. With this work, the aim is to remove this restriction, such that an mobile robot that is capable of traversing stairs will be able to explore an environment and map the space (with Simultaneous Localization And Mapping, for example) while simultaneously localizing the stairwells that exist in the environment within the map. With a map of the environment and estimated locations of the stairwells, the robot would be able to plan a path that traverses the stairs in order to explore the frontier at other elevations that were previously inaccessible. Some of these components already exist in some form, but an integrated solution for localizing and traversing stairs in the context of exploration has not yet been proposed. This work fills that need.

To be used for autonomous exploration, a stair detection system must be robust to differences in stairwell appearance, accuracy of detections and low false positive rate, indoor and outdoor capabilities, and most importantly real-time performance. Computational efficiency is particularly critical because of the demands of the other processes involved in exploration; a detector such as this should run in the background and inform the map without dominating the resources of the platform.

Our proposed system directly addresses the needs of an exploratory platform for solving the problem defined above. It is composed of a stairway detection module for extracting stair edge points in 3D from depth imagery and a stairway modeling module that aggregates many such detections into a single point cloud from which the stairway's dimensions and location are estimated (see Fig. 6). Modeling the stairway over many detections allows the system to form a complete model from many partial observations. We model the stairway as a single object: an inclined plane constrained by a bounding box, with stair edges lying in the plane. As new observations are added to the aggregated point cloud, the model is re-estimated, outliers are removed, and well-supported stair edges are used to infer the dimensions of each step.

We have deployed this system on an iRobot PackBot as well as a Turtlebot, both fitted with Microsoft Kinect depth sensors. Our system runs in real time and demonstrates robust and accurate performance in

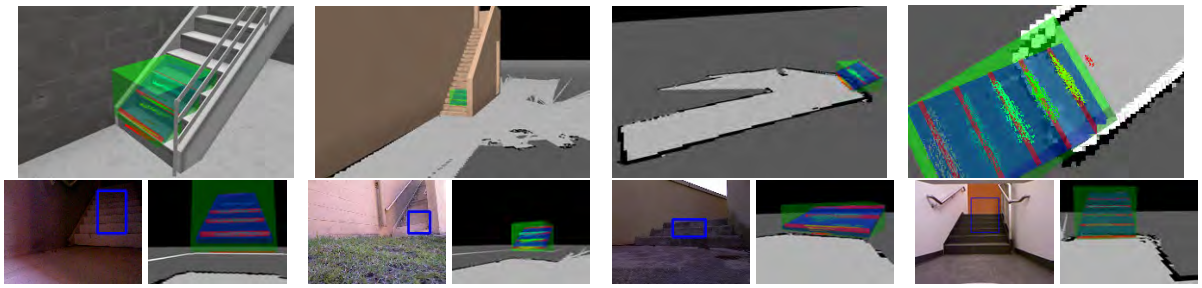

Figure 7: Results of several runs from our datasets: Building 1 Rear (first), Building 3 (second), Building 7 Exterior (third), and Davis Hall Rear (fourth).

both localization and parameter estimation for a wide variety of stairways. Our system has been tested extensively on data collected at a Military Operations in Urban Terrain (MOUT) site on all of the available stairway types at the site, as well as on numerous negative examples. There were over 10 stairway types throughout the site, both indoor and outdoor, of a variety of dimensions, and ranging from a few steps to a full flight. It has also been tested at a building at the SUNY at Buffalo (UB). These datasets consist of 9 recorded trials (7 and 2, respectively). Fig. 7 presents example results of our method on the real data.

- J. A. Delmerico, D. Baran, P. David, J. Ryde, and **J. J. Corso**. Ascending stairway modeling from dense depth imagery for traversability analysis. In *Proceedings of IEEE International Conference on Robotics and Automation*, 2013.
- J. A. Delmerico, **J. J. Corso**, D. Baran, P. David, and J. Ryde. Ascending stairway modeling: A first step toward autonomous multi-floor exploration. In *Proceedings of IEEE/RSJ Intelligent Robots and Systems (Video Proceedings)*, 2012.

Attributed object maps with object sensing while mapping.

An integral component of the semantic filtering project is the detection and recognition of scene components. In the past years of this project, we have emphasized macro scale objects, such as building facades. However, in this current year, we shifted our emphasis to consider small-scale objects that one may find inside a house or a building. We report on our methods for detection and reconstruction of such objects.

Appearance-based object detectors that produce bounding boxes are well developed and robust, with state of the art methods such as Discriminatively Trained Part Based Models [FGMR10] capable of achieving high accuracy. However, bounding boxes capture background and occlusions in addition to the object in question, so the actual object must be segmented from that bounding box in order to be isolated, manipulated, etc. Although bounding boxes are often roughly centered on the object they are detecting, this is not always the case, and non-convex objects that are centered may not actually occupy the geometric center of the bounding box (e.g. a torus). Additionally, a general class-level object detector will not make assumptions about object color or pose, or spatial layout if it is an RGB-D detector. Indeed, the minimal output we can expect from an object detector is a bounding box in the image and a class label, and our approach to automatic foreground object segmentation relies only on these inputs.

We seek to employ these foreground object segmentations in a bottom-up approach to 3D object reconstruction without an instance-level object model or class-level template so that arbitrary instances of an object class can be modeled in unknown environments. However, as with any computer vision system, there will be some noise and error in the segmentations: some visible parts of the object may be missing and some background or occlusions may be included. The challenge of this problem is the separation of object points from outliers using only the information contained in a collection of these segmentations.

Existing works build on established ideas such as background subtraction, geometric templates, table-top assumptions constraining pose and location of specific objects. We seek to relax these assumptions and avoid the aforementioned limitations in enabling foreground object segmentation for 3D object reconstruction. We have developed an approach that is general, extensible to arbitrary classes and scenarios, and can work online. A visual overview of our approach can be found in Fig. 8.

We want to leverage the existing work on appearance-based object detection in the literature. We therefore only assume that some other system provides object detections as bounding boxes in RGB-D images along with a class label. We make no assumptions about the specific appearance, geometry, or pose of a detected object. Instead, we learn a non-parametric model that captures the distribution of the foreground class in a histogram encoding location within the bounding box and object scale. The scale of the object is inferred from the apparent size of the object in the scene and its depth. We also learn a model of the spatial distribution and scale of the background, and use these models as priors for a Markov Random Field (MRF) that is used for two-class (foreground/background) labeling. We use spatial and color cues from the D and RGB channels of the RGB-D images to enforce segmentation boundaries when we perform energy minimization on the MRF to obtain automatic foreground segmentations.

For 3D object reconstruction, we combine multiple segmentations together, extracting the 3D points from each foreground segmentation and aggregating them into one point cloud. The resulting cloud will have many erroneous points from imperfect segmentations. However, regions in space that are repeatedly observed (the object) will have a high local point density relative to regions that are erroneously segmented as foreground in a few of the input images. Based on this intuition, we expect the density of the point cloud in the location of the object to be reinforced over multiple observations, allowing us to filter out occlusion and

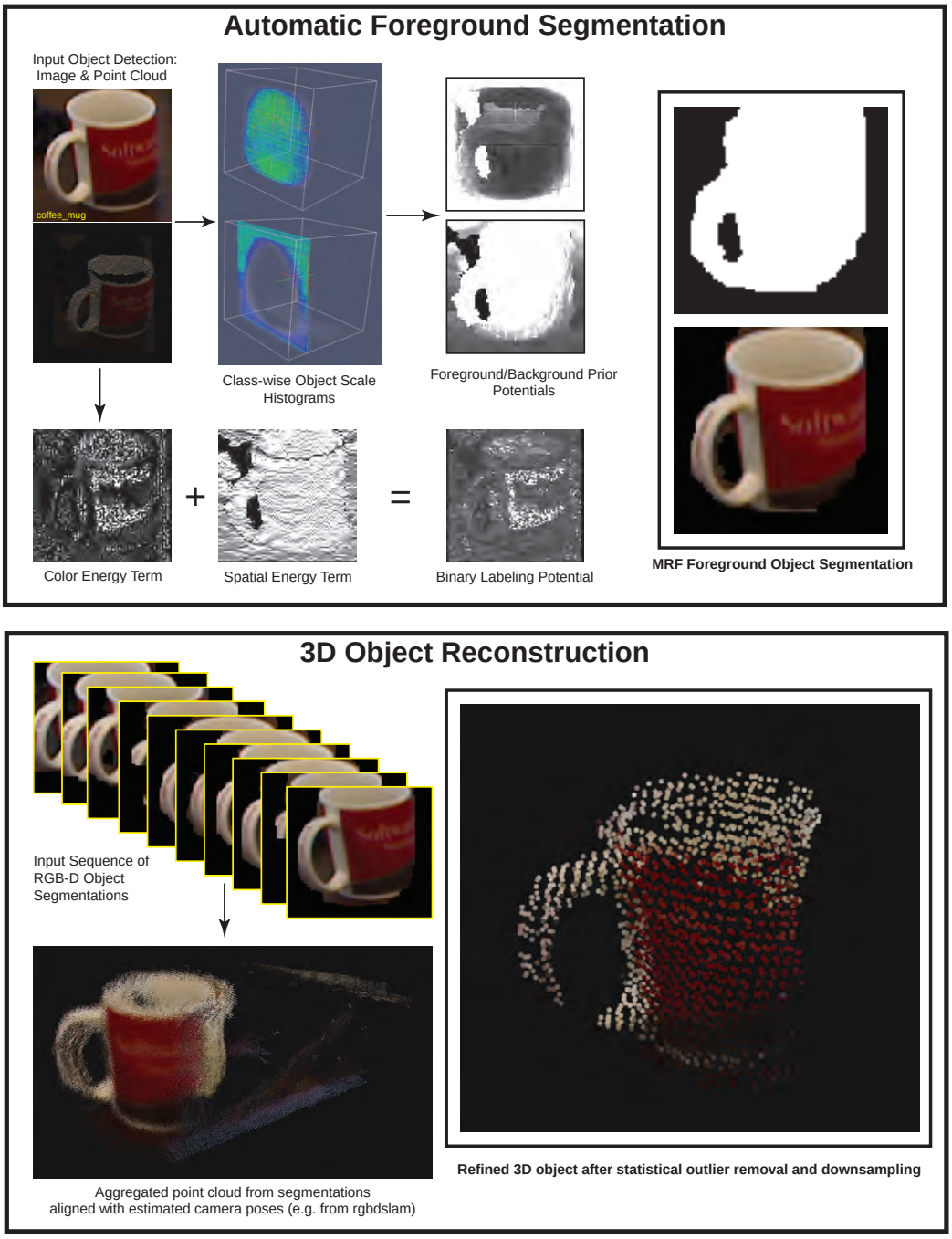

Figure 8: Method overview. Automatic segmentation is performed on an RGB-D bounding box input using a Markov Random Field. Multiple segmentations are aggregated in 3D and then the reconstruction is refined based on local point density.

background outliers based on their local point density. Our priors and constraints can be computed quickly from the image, and the segmentation routine is fast enough to be run in real time. Periodic refinement of the 3D reconstruction would allow this approach to be implemented online.

We make three primary contributions in this work:

- An algorithm for automatic foreground object segmentation from a bounding box detection using RGB-D cues in a Markov Random Field. Compatible with existing state of the art appearance-based detectors.
- A novel approach to modeling the likelihood of foreground and background classes within a bounding box based on relative location within the box in x and y and object scale inferred from its depth.
- A 3D object reconstruction algorithm that merges multiple partial observations. The segmentation algorithm runs fast enough to enable real time, online performance for this reconstruction approach.

- J. A. Delmerico. *Attributed Object Maps: Descriptive Object Models as High-level Semantic Features for Mobile Robotics*. PhD thesis, State University of New York at Buffalo, 2013.

Building Facade Detection with Layered Graphical Models.

One semantic inference example is to segment and model building facades from stereo imagery (Figure 9). Obtaining accurate models of the facades in a scene can assist a mobile platform in localization and guidance.

The discriminative model developed for the facade detection phase of this project is leveraged to perform classification of pixels in the image into our two classes: building and background. The method depends on having a camera and 2D laser scanner registered such that the laser scan points that fall within the field of view of the camera can be isolated to individual pixels of the image. The facade classifier extracts only those points from the scan that both fall within the cameras field of view and are classified as being on building in the scene. This filtered scan is then passed to mapping and localization processes on the robot.

A discriminative model is generated from an extension of the Boosting on Multilevel Aggregates (BMA) method [**Corso**08] that includes stereo features. Boosting on Multilevel Aggregates uses hierarchical aggregate regions coarsened from the image based on pixel affinities, as well as a variety of high-level features that can be computed from them, to learn a model within an AdaBoost two- or multi-class discriminative modeling framework. The multilevel aggregates exploit the propensity of these coarsened regions to adhere to object boundaries, which in addition to the expanded feature set, offer less polluted statistics than standard patch-based features, which may violate those boundaries. Since many mobile robot platforms are equipped with stereo cameras, and can thus compute a disparity map for their field of view, our approach of using statistical features of the disparity map is a natural extension of the BMA approach given our intended platform. Since buildings tend to have planar surfaces on their exteriors, the stereo features are used to exploit the property that planes can be represented as linear functions in disparity space and thus have constant spatial gradients.

In order to associate each building pixel with a particular facade, a set of candidate planes from which to infer the best fit is generated. Sampling the image and performing Principal Component Analysis (PCA) on each local neighborhood to approximate the local surface normal at the sampled points generate these. Those points are then clustered by iteratively using Random Sample Consensus (RANSAC) to find subsets that fit the same plane model and have similar local normal orientations. From these clusters of points, the parameters of the primary planes are estimated.

Finally, both sources of information are incorporated into a two-layer Markov random field with an Ising model at the middle level on the plane detections. The high-level representation is a Potts model, where each hidden variable represents the labeling of the associated pixel with one of the candidate planes, or with no plane if it is not part of a building. Graph cuts energy minimization method to compute minimum energy labelings for both levels of the MRF model.

The primary contributions of this work are a novel approach to discriminative modeling for building facade detection that leverages stereo data, a top-down plane fitting procedure on the disparity map, and a novel Markov random field model for fusing the appearance model from the discriminative classification and the geometric model from the plane fitting step to produce a facade segmentation of a single-view stereo image.

This system has been implemented and was deployed on a Packbot during Jeffrey Delmericos summer internship at ARL in 2011, and tested more extensively during field tests with ARL at Camp Lejeune in October 2011. A version of the faade classifier based on standard Adaboost (and not the stereo classifier) was used in order to achieve real-time performance. The scan classification performed very well on a variety of buildings and settings. The mapping aspect of this project has not yet been as successful, resulting in maps

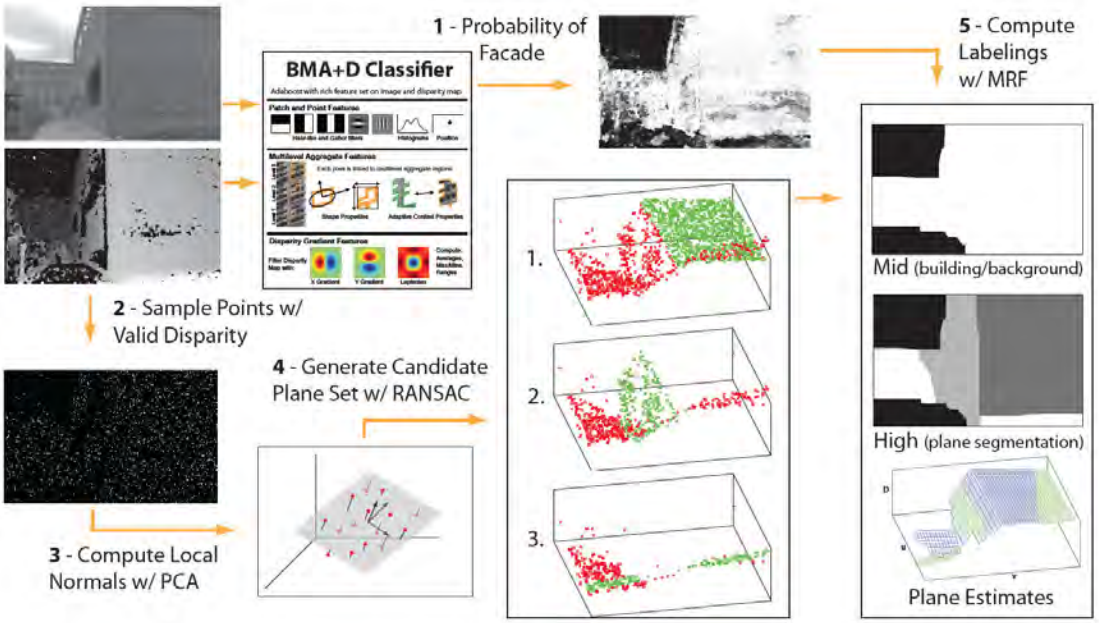

Figure 9: Overview of the two-layer model for building facade detection to support semantic filtering.

that are adequate, but not obviously more accurate or noise-free than those produced without filtering. This may be due at least in part to the narrow field of view of the camera; a laser scanner typically produces a scan with a 200+ degrees field of view, but with filtering this is reduced to approximately 60 degrees, making the scan matching procedure much more difficult for the mapping and localization routine.

In addition, we have presented a poster at the August 2011 NSF workshop on Frontiers in Computer Vision, which is available at `http://www.cse.buffalo.edu/~jcorso/pubs/jcorso_fcv_poster.pdf`.

- J. A. Delmerico, P. David, and **J. J. Corso**. Building facade detection, segmentation, and parameter estimation for mobile robot localization and guidance. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2011.

Parts-based Approach to Semantic Segmentation/Pixel Labeling

This material is partially supported by ARL/ARO W911NF-11-1-0090.

Semantic pixel labeling in images and videos. Our work in semantic pixel labeling into background and foreground classes has continued and resulted in new findings this year in support of full scene semantic segmentation for improved guidance by the more rich semantics.

Scene understanding remains a significant challenge in the computer vision community. The visual psychophysics literature has demonstrated the importance of interdependence among parts of the scene. Yet, the majority of methods in computer vision remain local. Pictorial structures have arisen as a fundamental parts-based model for some vision problems, such as articulated object detection. However, the form of classical pictorial structures limits their applicability for global problems, such as semantic pixel labeling. In this paper, we propose an extension of the pictorial structures approach, called pixel-support parts-sparse pictorial structures, or PS3, to overcome this limitation. Our model extends the classical form in two ways: first, it defines parts directly based on pixel-support rather than in a parametric form, and second, it specifies a space of plausible parts-based scene models and permits one to be used for inference on any given image. PS3 makes strides toward unifying object-level and pixel-level modeling of scene elements. In this report, we implement the first half of our model and rely upon external knowledge to provide an initial graph structure for a given image. Our experimental results (Figure 10 on benchmark datasets demonstrate the capability of this new parts-based view of scene modeling.

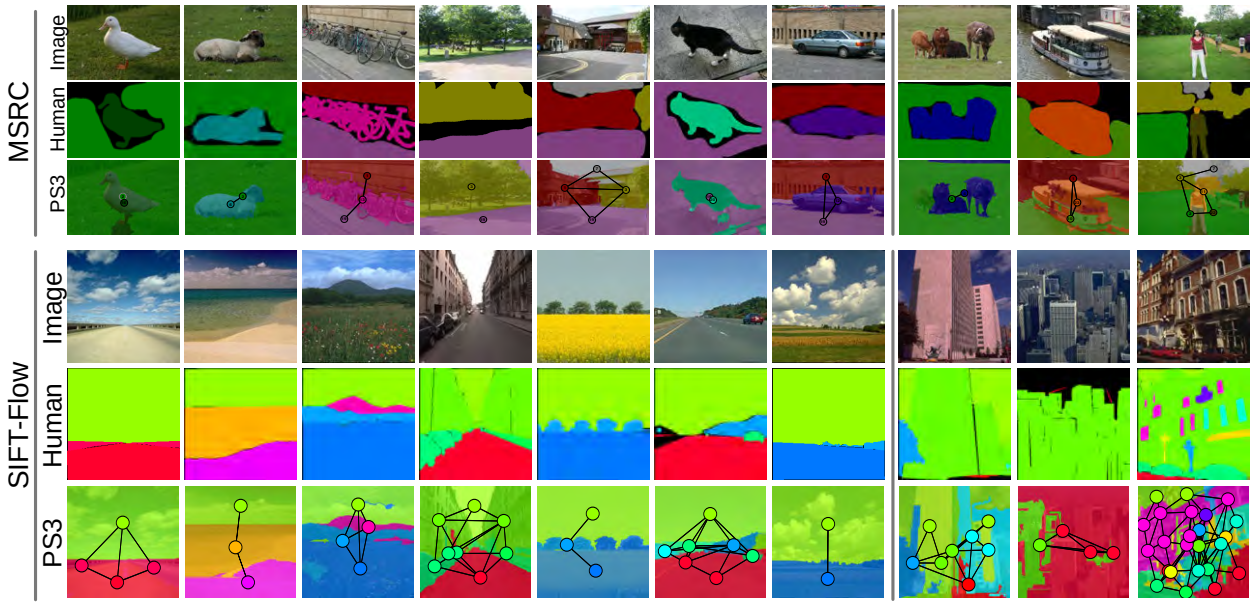

Figure 10: Visual semantic pixel labeling results with our new PS3 model on the two data sets. Each columns shows an example in three rows: (1) original image, (2) human gold standard, and (3) our PS3 result overlaid upon the image. We have also rendered the graph structure on top of the image. The results on the right side of the figure show some of the worst examples of our performance.

- **J. J. Corso**. Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures.

Pattern Recognition Letters: Special Issue on Scene Understanding and Behavior Analysis, 34(7):762–769, 2013. Early version appears as arXiv.org tech report 1108.4079v1.

- A. Y. C. Chen and **J. J. Corso**. Temporally consistent multi-class video-object segmentation with the video graph-shifts algorithm. In *Proceedings of the 2011 IEEE Workshop on Motion and Video Computing*, 2011.

- A. Y. C. Chen and **J. J. Corso**. Propagating multi-class pixel labels throughout video frames. In *Proceedings of Western New York Image Processing Workshop*, 2010.

Ascending Stairway Detection and Traversability Estimation

The goal is to use geometric cues from depth data to detect ascending stairwells and operate at the frame rate of the sensor when deployed on a mobile platform. Detection of stairwells by mobile robots will enable multi-floor exploration and mapping for those platforms capable of stair traversal.

Autonomous mobile robots have traditionally been restricted to exploring single floors of a building or outdoor areas that are free of abrupt elevation changes such as curbs and stairs. With this work, the aim is to remove this restriction, such that an mobile robot that is capable of traversing stairs will be able to explore an environment and map the space (with Simultaneous Localization And Mapping, for example) while simultaneously localizing the stairwells that exist in the environment within the map. With a map of the environment and estimated locations of the stairwells, the robot would be able to plan a path that traverses the stairs in order to explore the frontier at other elevations that were previously inaccessible. Some of these components already exist in some form, but an integrated solution for localizing and traversing stairs in the context of exploration has not yet been proposed. This work fills that need.

To be used for autonomous exploration, a stair detection system must be robust to differences in stairwell appearance, accuracy of detections and low false positive rate, indoor and outdoor capabilities, and most importantly real-time performance. Computational efficiency is particularly critical because of the demands of the other processes involved in exploration; a detector such as this should run in the background and inform the map without dominating the resources of the platform. With these constraints driving our design, the proposed system exploits the geo- metric properties of stairs that exist in depth images. It applies a number of image processing techniques (Canny edge detection, probabilistic Hough transform) to detect lines in the depth image representing discontinuities; stairwells will exhibit multiple parallel discontinuities. These candidate lines are filtered to isolate a set that are parallel (or nearly so) and localized as a group in the image. Finally, a plane is fit to the detected stair edge lines to confirm that the detected stair edges lie on an inclined plane that at a traversable angle. The detector has been deployed on an iRobot Packbot equipped with a Microsoft Kinect depth sensor in a variety of natural environments and has proven to succeed in addressing the needs of a robotic system for autonomous exploration; it is very computationally lightweight, robust to stair size, appearance, and viewing angle, highly accurate, and (in principle) applicable to depth imagery captured both indoor and out.

This work is currently underway and accepted as to the Video Proceedings of IROS 2012. We will include more detail on it once the full publication is ready and released.

Figure 12: A montage of entries in the action bank, 36 of the 205 in the bank. Each entry in the bank is a single template video example. The columns depict different types of actions, e.g., a baseball pitcher, boxing, etc. and the rows indicate different examples for that action. Examples are selected to roughly sample the action's variation in viewpoint and time (but each is a different video/scene, i.e., this is not a multiview requirement).

Action Bank: A High-Level Representation of Activity

This material is partially supported by ARL/ARO W911NF-11-1-0090.

Human motion and activity is extremely complex; automatically inferring activity from video in a robust manner that would lead to a rich high-level understanding of video remains a challenge despite the great energy the vision community has invested in it. The most promising current approaches are primarily based on low- and mid-level features such as local space-time features [Lap05], dense point trajectories [WKSL11], and dense 3D gradient histograms [KMS08] to name a few; these methods have demonstrated capability on realistic data sets like UCF Sports [RAS08]. But, they are, by nature, limited in the amount of motion semantics they can capture being strictly low-level, which often yields a representation with inadequate discriminative power for larger, more complex data sets. For example, on the 50-class UCF50 data set [UCF], the HOG/HOF method [Lap05, WUK+09] achieves 47.9% accuracy (as reported in [KJG+11]) whereas it achieves 85.6% on the smaller 9-class UCF Sports data set [RAS08]. Other methods that seek a more semantically rich and discriminative representation have focused on object and scene semantics [HYL+07] or human pose, e.g., [ABS07, RF03], which itself is challenging and unsolved.

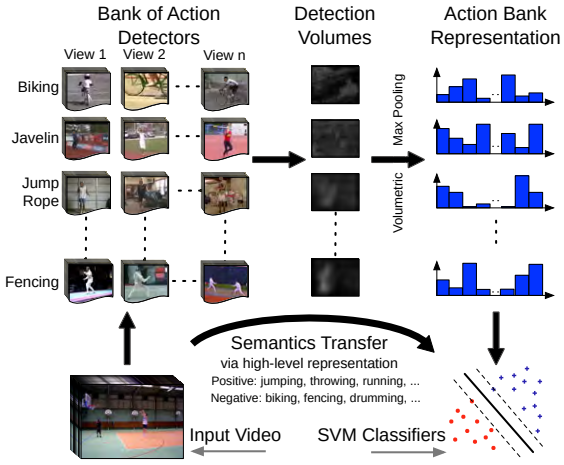

Figure 11: Action bank is a high-level representation for video activity recognition.

In this work, we propose a new high-level representation of human action in video that we call Action Bank. Inspired by the Object Bank method [LSXFF10], action bank explores how a large set of action detectors, which ultimately act like the bases of a high-dimensional "action-space," combined with a simple linear classifier can form the basis of a semantically-

rich representation for activity recognition and other video understanding challenges (Figure 11 shows an overview). The individual action detectors in the action bank are based on an adaptation of the recent action spotting framework [DSCW10] and hence template-based; despite the great amount of research on action recognition, few methods are available that localize action in the video as a detector must. Individual detectors in the bank capture example actions, such as "running-left" and "biking-away," and are run at multiple scales over the input video (many examples of detectors in action bank are shown in Figure 12). The outputs of detectors are transformed into a feature vector by volumetric max-pooling. Although the resulting vector is high-dimensional, we test an SVM classifier that is able to enforce sparsity among its representation, in a manner similar to object bank.

Although there has been some work on mid- and high-level representations for video recognition and retrieval [HYL+07], to the best of our knowledge it has exclusively been focused on object and scene-level semantics, such as face detection. Our work hence breaks new ground in establishing a high-level representation built atop individual action detectors. We show that this high-level representation of human activity is capable of being the basis of a powerful activity recognition method, achieving better than state-of-the-art accuracies on every major activity recognition benchmark attempted, including 98.2% on KTH [SLC04], 95.0% on UCF Sports [RAS08], 57.9% on the full UCF50 [UCF], and 26.9% on HMDB51 [KJG+11]. Furthermore, action bank also transfers the semantics of the individual action detectors through to the final classifier (Figure 13).

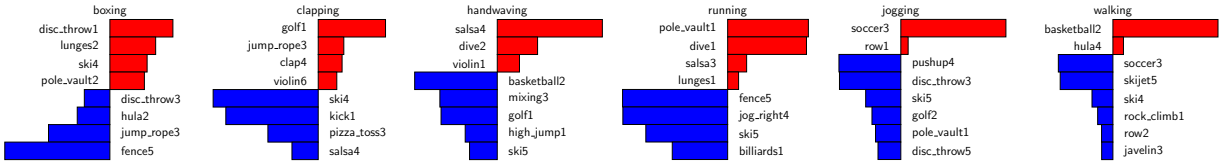

Figure 13: Relative contribution of the dominant positive and negative bank entries for each one-vs-all SVM on the KTH data set. The action class is named at the top of each bar-chart; red (blue) bars are positive (negative) values in the SVM vector. The number on bank entry names denotes which example in the bank (recall that each action in the bank has 3–6 different examples). Note the frequent semantically meaningful entries; for example, "clapping" incorporates a "clap" bank entry and "running" has a "jog" bank entry in its negative set.

- S. Sadanand and **J. J. Corso**. Action bank: A high-level representation of activity in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

LIBSVX: Supervoxels and Streaming Video Segmentation

This material is partially supported by ARL/ARO W911NF-11-1-0090.

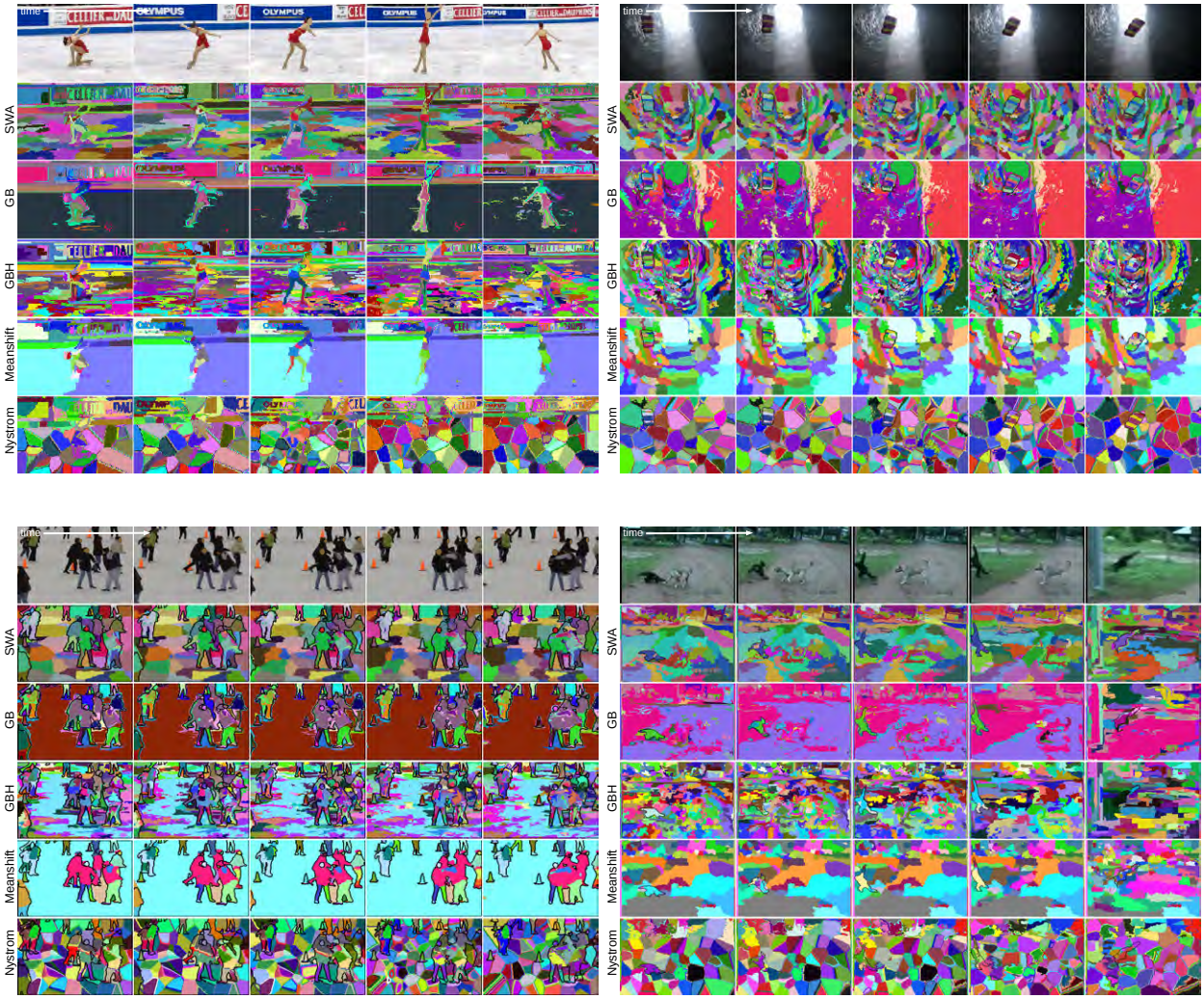

Figure 14: Visual comparative results of the five methods on four videos (with roughly 500 supervoxels); top-left is from Gatech, bottom-left is from Chen, and right-two are from SegTrack. A black line is drawn to represent human-drawn boundaries in those videos that have been annotated; note for the SegTrack data set only one object has a boundary and for the Chen data set many regions have boundaries. Each supervoxel is rendered with its distinct color and these are maintained over time. Faces have been redacted for presentation (the original videos were processed when computing the segmentations). We recommend viewing these images zoomed on an electronic display.

Images have many pixels; videos have more. It has thus become standard practice to first preprocess images and videos into more tractable sets by either extraction of salient points [SM97] or oversegmentation into superpixels [RM03]. The preprocessing output data—salient points or superpixels—are more perceptually

meaningful than raw pixels, which are merely a consequence of digital sampling [RM03]. However, the same practice does not entirely exist in video analysis. Although many methods do indeed initially extract salient points or dense trajectories, e.g., [Lap05], few methods we are aware of rely on a supervoxel segmentation, which is the video analog to a superpixel segmentation. In fact, those papers that do preprocess video tend to rely on a per-frame superpixel segmentation, e.g., [LKG11], or use a full-video segmentation, e.g., [GKHE10].

We have performed a thorough comparative evaluation of five supervoxel methods [20]; note that none of these methods had been proposed intrinsically as a supervoxel method, but each is either sufficiently general to serve as one or has been adapted to serve as one. The five methods we choose—segmentation by weighted aggregation (SWA) [SBB00, SGS$^+$06, **Corso**SD$^+$08], graph-based (GB) [FH04], hierarchical graph-based (GBH) [GKHE10], mean shift [PD07], and Nyström normalized cuts [FBCM04, SM00, FBM01]—broadly sample the methodology-space, and are intentionally selected to best analyze methods with differing qualities for supervoxel segmentation. For example, both the SWA and the Nyström method use the normalized cut criterion as the underlying objective function, but SWA minimizes it hierarchically whereas Nyström does not. Similarly, there are two graph-based methods that optimize the same function, but one is subsequently hierarchical (GBH). We note a similar selection of segmentation methods have been used in the (2D) image boundary comparative study [AMFM11].

We have conducted a benchmark evaluation of these five supervoxels; the benchmark proposes a suite of desiderata that determine a good supervoxel method. Examples of the criteria include boundary recall, explained variation, and undersegmentation error, all in 3D. Our thorough evaluation of five supervoxel methods on four 3D volumetric performance metrics designed to evaluate supervoxel desiderata. We use the Chen Xiph.org annotated video data set as one of three core data sets within the evaluation (Samples from the data sets segmented under all five methods are shown in Figure 14). We have selected videos of different qualities to show in this figure.

The visual results convey the overall findings we have observed in the quantitative experiments (for example, see Figure 15 for a 3D boundary recall comparison plot). Namely, two of the hierarchical methods (GBH and SWA) perform better than the others at preserving object boundaries. The Nyström supervoxels are the most compact and regular in shape and the SWA supervoxels observe a similar compactness but seem to adapt to object boundaries better (recall that SWA and Nyström are both normalized cut solvers). It seems evident that the main distinction behind the better performance of GBH and SWA is the way in which they both compute the hierarchical segmentation. Although the details differ, the common feature among the two methods is that during the hierarchical computation, coarse-level aggregate features replace or modulate fine-level individual features. None of the other three approaches use any coarse-level features.

Finally, we report on our extension of these higher performing hierarchical graph-based methods to handle arbitrary long videos [21]. The use of video segmentation as an early processing step in video analysis lags behind the use of image segmentation for image analysis, despite many available video segmentation methods. A major reason for this lag is simply that videos are an order of magnitude bigger than images; yet most methods require all voxels in the video to be loaded into memory, which is clearly prohibitive for even medium length videos. We address this limitation by proposing an approximation framework for streaming hierarchical video segmentation motivated by *data stream* algorithms: each video frame is processed only once and does not change the segmentation of previous frames. We implement the graph-based hierarchical segmentation method within our streaming framework; our method is the first streaming hierarchical video segmentation method proposed.

For a given video $\mathcal{V}$, consider an objective function or criterion $E(\cdot|\cdot)$ to obtain the hierarchical segmentation

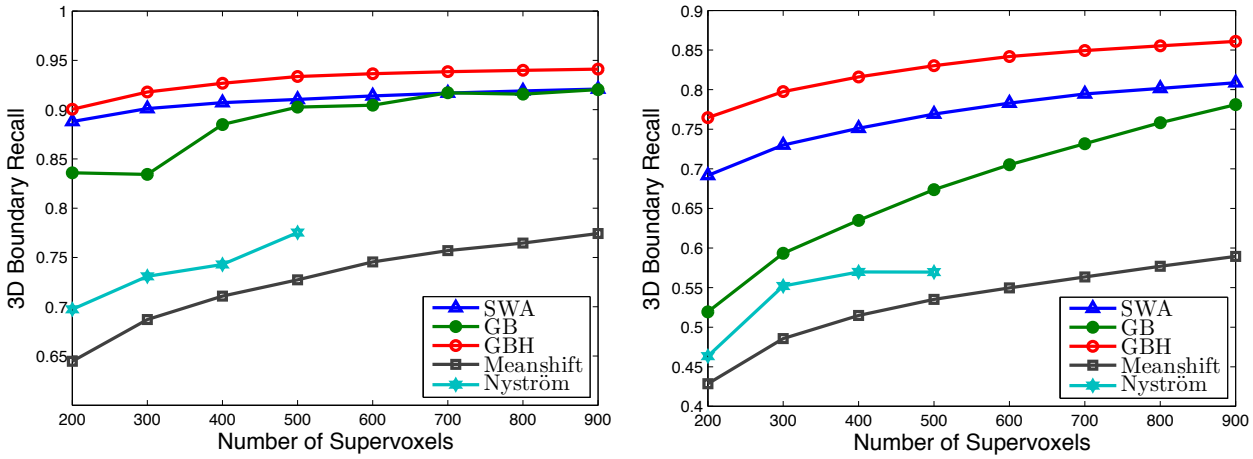

Figure 15: 3D boundary recall vs. number of supervoxels. Left: the results on SegTrack data set. Right: the results on Chen's data set.

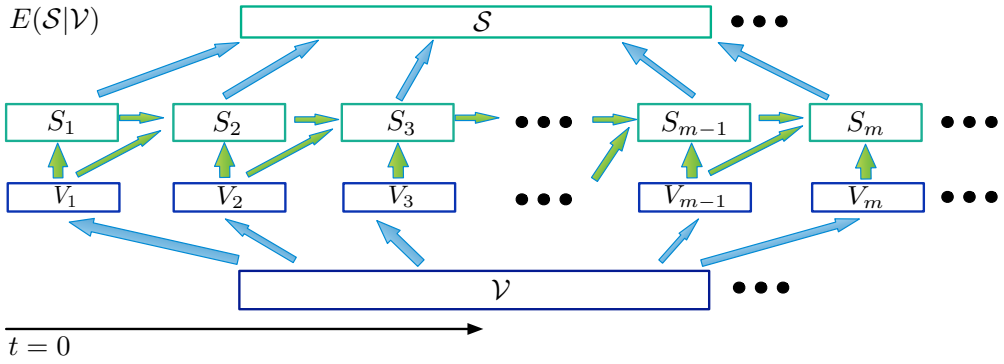

Figure 16: Framework of streaming hierarchical video segmentation.

result $\mathcal{S}$ by minimizing:

$$\mathcal{S}^* = \operatorname*{argmin}_{\mathcal{S}} E(\mathcal{S}|\mathcal{V}) \ . \tag{3}$$

Concretely, let Λ^2 denote the 2D pixel lattice. We think of a video as a function on space-time lattice $\Gamma \doteq \Lambda^2 \times \mathbb{Z}$ mapping to color space $\mathbb{R}^3$, s.t. $\mathcal{V}\colon \Gamma \to \mathbb{R}^3$. The hierarchical segmentation results in h layers of individual segmentations $\mathcal{S} \doteq \{S^1, S^2, \dots, S^h\}$ where each layer S^i is a set of segments $\{s_1, s_2, \dots\}$ such that $s_j \subset \Gamma$, $\cup_j s_j = \Gamma$, and $s_i \cap s_j = \varnothing$ for all pairs of segments. Each layer in the hierarchy gives a full segmentation and the hierarchy itself defines a forest of trees (segments only have one parent). There are several methods, e.g., [GKHE10, Par08], for pursuing this hierarchical segmentation. However, we are not aware of any method that has an explicit objective function to optimize; they are all based on some criteria to pursue hierarchical segmentation results. We consider them as greedy/gradient descent-like methods to approximate the global or local minimal of an implicit objective function $E(\mathcal{S}|\mathcal{V})$.

Our major contribution is an approximation framework for evaluating (3), displayed in Figure 16 graphically. The approximation framework is based on a pair of Markov assumptions, one temporally, and one

hierarchically that, together, all of the streaming hierarchical segmentation of the video. The temporal approximation uses a stream pointer, t, to index into the video; t may only more forward in time (from data streams). For every $t + \tau$ frames, we process the video segmentation at the pixel level (needing on only the current $t + \tau$ and the previous $t - \tau$ through $t - 1$ frames in memory. The hierarchical approximation naturally requires only the next finer level segmentation available for the current level segmentation.

In this framework, we have implemented the graph-based minimum spanning tree method from [FH04], which was extended to video by [GKHE10] but not streaming hierarchically. We have conducted extensive quantitative experiments for our new method, using the benchmark from our CVPR 2012 paper [20]. Figure 17 presents a sample of them demonstrating strong performance of the approximation results in terms segmentation performance with respect to original method that is not streaming (and hences requires quite more memory) and the other existing streaming methods (which do not perform as well on the metrics).

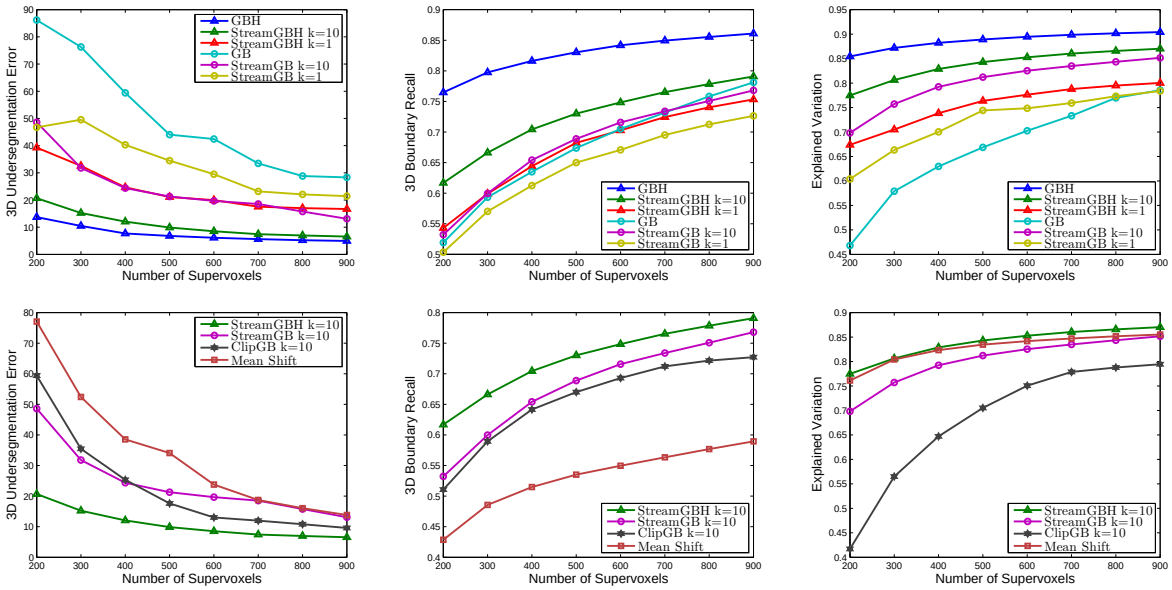

Figure 17: Quantitative experiments on the benchmark data set. Left: 3D Undersegmentation Error. Middle: 3D Boundary Recall. Right: Explained Variation. Top Row: performance of Streaming GB/GBH with different k against full-video versions. Bottom Row: comparison against streaming methods.

The full LIBSVX software distribution is published at `http://www.cse.buffalo.edu/~jcorso/r/supervoxels` and is freely usable for follow-on work.

- C. Xu and **J. J. Corso**. Evaluation of super-voxel methods for early video processing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

- C. Xu, C. Xiong, and **J. J. Corso**. Streaming hierarchical video segmentation. In *Proceedings of European Conference on Computer Vision*, 2012.

Max-Margin and Random Forest Based Metric Learning

This material is partially supported by ARL/ARO W911NF-11-1-0090.

Metric learning makes it plausible to learn semantically meaningful distances for complex distributions of data using label or pairwise constraint information. However, to date, most metric learning methods are based on a single Mahalanobis metric, which cannot handle heterogeneous data well. Those that learn multiple metrics throughout the feature space have demonstrated superior accuracy, but at a severe cost to computational efficiency. Here, we adopt a new angle on the metric learning problem and learn a single metric that is able to implicitly adapt its distance function throughout the feature space. This metric adaptation is accomplished by using a random forest-based classifier to underpin the distance function and incorporate both absolute pairwise position and standard relative position into the representation.

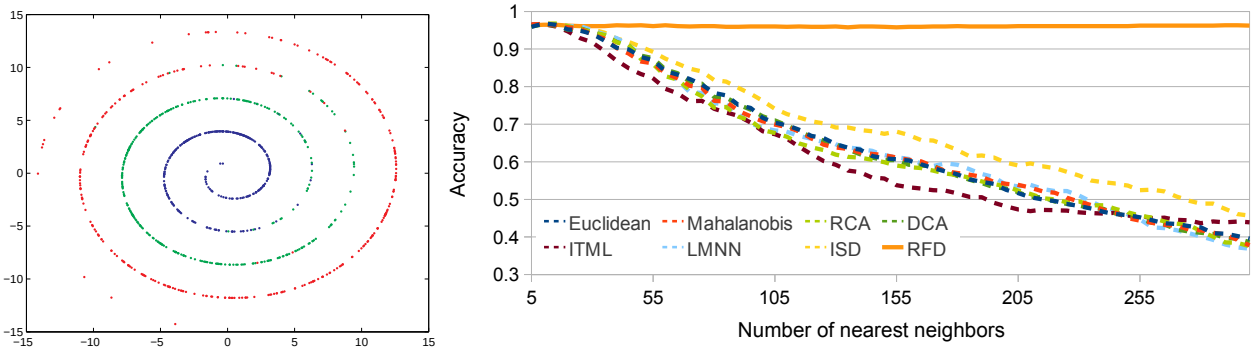

Figure 18: An example using a classic swiss roll data set comparing both global and position-specific Mahalanobis-based methods with our proposed method, RFD. All methods, including the baseline Euclidean, perform well at low k-values due to local linearity. However, as k increases and the global nonlinearity of the data becomes important, the monolithic methods' inability to incorporate position information causes their performance to degrade until it is little better than chance. The position-specific ISD method performs somewhat better, but even with a Mahalanobis matrix at every point it is unable to capture the globally nonlinear relations between points. Our method, by comparison, shows no degradation as k increases. (3 classes, 900 samples, validated using k-nearest neighbor classification, with varying k)

We have implemented and tested our method against state of the art global and multi-metric methods on a variety of data sets. An example result is presented in Figure 18. Overall, the proposed method outperforms both types of method in terms of accuracy (consistently ranked first) and is an order of magnitude faster than state of the art multi-metric methods (16x faster in the worst case).

Our second finding on the metric learning side uses a max-margin objective function to learn the metric. Efficient learning of an appropriate distance metric is an increasingly important problem in machine learning. However, current methods are limited by scalability issues or are unsuited to use with general similarity/dissimilarity constraints. We have proposed an efficient metric learning method based on the max-margin framework with pairwise constraints that has strong generalization guarantee. First, we reformulate the max-margin metric learning problem as a structured support vector machine which we can optimize in linear time via a cutting-plane method. Second, we propose an approximation method for our kernelized extension based on match pursuit algorithm that allows linear-time training. We find our method

to be comparable to or better than state of the art metric learning techniques at a number of machine learning and computer vision classification tasks. This work won the Best Paper prize at ECDM [17].

- C. Xiong, D. Johnson, R. Xu, and **J. J. Corso**. Random forests for metric learning with implicit pairwise position dependence. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.
- C. Xiong, D. Johnson, and **J. J. Corso**. Efficient max-margin metric learning. In *Proceedings of European Conference on Data Mining*, 2012. **Winner of Best Paper Award at ECDM 2012.**.

3 Scientific Significance

The majority of the work in semantic based mobile robot sensing focuses on a priori objects and environments in a given map. The work in this project continues to break new ground in dynamically perceiving semantic properties of the scene and incorporating them into the control loop of the platform. This is a significant shift in the way visual inference is leveraged for mobile robot applications. Our work has, for example, impacted the visual inference routines in a mobile robot application at the ARL.

4 Summary of Accomplishments

We accomplished numerous technology advances, including novel methods for visual inference in video, innovative mathematical formulations for assessing traversability of stairways and for mutual localization of multiple robot platforms, and for evaluating guidance by semantics on real-world mobile robot data.

5 Collaborations and Leveraged Funding

We continue a steady and fruitful collaboration with Army Research Labs Adelphi, MD. In particularly, our team is in close contact with Dr. Philip David, Dr. Stuart Young, Dr. David Baran and Dr. Cynthia Pierce. A weekly telecom has been held between our team and Dr. Philip David for parts of the project execution; Jeffrey Delmerico, from Buffalo, has interned during the summers of 2010, 2011 and 2012 at ARL both cases resulting in publication output and technology transfer; and David Johnson, from Buffalo, is currently visiting the ARL for an internship (he also spent the summer of 2014 at the ARL as an intern).

The ARO DURIP grant the PI received has enabled him to stand up a mobile robot testbed within his lab at the University. This has already made it easier to facilitate in-house testing and faster transition to ARL.

The PI has been involved in numerous national research programs, including the DARPA CSSG, DARPA MINDSEYE, IARPA ALADDIN, NIH, and also has other support from the Army Research Office. In all of these programs, he is collaborating with many members of the research community. For example, on the Army Research Office project, he is collaborating with Dr. Philip David at ARL to incorporate the PI's semantic image labeling methods into vision-based robot guidance algorithms. In his DARPA CSSG project, he is collaborating with Prof. Matthias Kolsch at the Naval Postgraduate School to incorporate the object detection methods developed during the Summer of Code 2010 into a comprehensive tool for the intelligence community. All in all, the PI has demonstrated exceptional versatility in working with other scientists in varying capacities.

6 Technology Transfer

This list is continuous from the project inception.

Semantic Filtering. This system has been implemented and was deployed on a Packbot during Jeffrey Delmericos summer internship at ARL in 2011, and tested more extensively during field tests with ARL at Camp Lejeune in October 2011. More data acquisition was performed at Camp Lejeune in September 2012.

Stairway Detection. This system has been implemented and was deployed on a Packbot during Jeffrey Delmericos summer internship at ARL in 2011, and tested more extensively during field tests with ARL

at Camp Lejeune in October 2011 and September 2012. The detector exhibited very high accuracy and a negligible false-positive rate. The assessment routines also demonstrated quantitatively high accuracy for traversability (e.g., within a few degrees error incline).

Action Bank. The human activity recognition system developed in 2011 year, Action Bank, has received considerable commercial interest. The university has filed a patent in Dec. 2012 on it.

7 Full Publication List

We summarize all of the publications that were in part or full supported by this award; this is an all-inclusive list from the project's start.

[1] C. Xu, S.-H. Hsieh, C. Xiong, and **J. J. Corso**. Can humans fly? Action understanding with multiple classes of actors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[2] J. Lu, R. Xu, and **J. J. Corso**. Human action segmentation with hierarchical supervoxel consistency. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] W. Chen, C. Xiong, R. Xu, and **J. J. Corso**. Actionness ranking with lattice conditional ordinal random fields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] V. Dhiman, A. Kundu, F. Dellaert, and **J. J. Corso**. Modern MAP inference methods for accurate and faster occupancy grid mapping on higher order factor graphs. In *Proceedings of International Conference on Robotics and Automation*, 2014.

[5] C. Xiong, D. M. Johnson, and **J. J. Corso**. Active clustering with model-based uncertainty reduction. Technical Report 1402.1783, arXiv, 2014.

[6] P. Das, C. Xu, R. F. Doell, and **J. J. Corso**. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[7] J. A. Delmerico, D. Baran, P. David, J. Ryde, and **J. J. Corso**. Ascending stairway modeling from dense depth imagery for traversability analysis. In *Proceedings of IEEE International Conference on Robotics and Automation*, 2013.

[8] V. Dhiman, J. Ryde, and **J. J. Corso**. Mutual localization: Two camera relative 6-dof pose estimation from reciprocal fiducial observation. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2013.

[9] J. A. Delmerico, **J. J. Corso**, D. Baran, and P. David. Towards autonmous multi-floor exploration: Ascending stairway localization and modeling. Technical Report ARL-TR-6381, Army Research Laboratory, 2013.

[10] J. A. Delmerico. *Attributed Object Maps: Descriptive Object Models as High-level Semantic Features for Mobile Robotics*. PhD thesis, State University of New York at Buffalo, 2013.

[11] **J. J. Corso**. Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures. *Pattern Recognition Letters: Special Issue on Scene Understanding and Behavior Analysis*, 34(7):762–769, 2013. Early version appears as arXiv.org tech report 1108.4079v1.

[12] J. A. Delmerico, **J. J. Corso**, D. Baran, P. David, and J. Ryde. Ascending stairway modeling: A first step toward autonomous multi-floor exploration. In *Proceedings of IEEE/RSJ Intelligent Robots and Systems (Video Proceedings)*, 2012.

[13] J. Ryde and **J. J. Corso**. Fast voxel maps with counting bloom filters. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2012.

[14] S. Sadanand and **J. J. Corso**. Action bank: A high-level representation of activity in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] C. Xiong and **J. J. Corso**. Coaction discovery: Segmentation of common actions across multiple videos. In *Proceedings of Multimedia Data Mining Workshop in Conjunction with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (MDMKDD)*, 2012.

[16] C. Xiong, D. Johnson, and **J. J. Corso**. Spectral active clustering via purification of the k-nearest neighbor graph. In *Proceedings of European Conference on Data Mining*, 2012.

[17] C. Xiong, D. Johnson, and **J. J. Corso**. Efficient max-margin metric learning. In *Proceedings of European Conference on Data Mining*, 2012. **Winner of Best Paper Award at ECDM 2012.**.

[18] C. Xiong, D. Johnson, R. Xu, and **J. J. Corso**. Random forests for metric learning with implicit pairwise position dependence. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[19] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and **J. J. Corso**. Combining skeletal pose with local motion for human activity recognition. In *Proceedings of VII Conference on Articulated Motion and Deformable Objects*, 2012.

[20] C. Xu and **J. J. Corso**. Evaluation of super-voxel methods for early video processing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[21] C. Xu, C. Xiong, and **J. J. Corso**. Streaming hierarchical video segmentation. In *Proceedings of European Conference on Computer Vision*, 2012.

[22] J. A. Delmerico, P. David, and **J. J. Corso**. Building facade detection, segmentation, and parameter estimation for mobile robot localization and guidance. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2011.

References

[ABS07] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[AMFM11] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

[BG97] M. Betke and L. Gurvits. Mobile robot localization using landmarks. *Robotics and Automation, IEEE Transactions on*, 13(2):251–263, 1997.

[BRL+11] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(9):1820–1833, September 2011.

[C**Corso**10] A. Y. C. Chen and **J. J. Corso**. Propagating multi-class pixel labels throughout video frames. In *Proceedings of Western New York Image Processing Workshop*, 2010.

[C**Corso**11] A. Y. C. Chen and **J. J. Corso**. Temporally consistent multi-class video-object segmentation with the video graph-shifts algorithm. In *Proceedings of the 2011 IEEE Workshop on Motion and Video Computing*, 2011.

[**Corso**08] **J. J. Corso**. Discriminative Modeling by Boosting on Multilevel Aggregates. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[**Corso**13] **J. J. Corso**. Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures. *Pattern Recognition Letters: Special Issue on Scene Understanding and Behavior Analysis*, 34(7):762–769, 2013. Early version appears as arXiv.org tech report 1108.4079v1.

[**Corso**SD+08] **J. J. Corso**, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille. Efficient Multilevel Brain Tumor Segmentation with Integrated Bayesian Model Classification. *IEEE Transactions on Medical Imaging*, 27(5):629–640, 2008.

[CSF+12] M. Cognetti, P. Stegagno, A. Franchi, G. Oriolo, and H.H. Bulthoff. 3-D mutual localization with anonymous bearing measurements. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 791 –798, may 2012.

[CXX**Corso**14] W. Chen, C. Xiong, R. Xu, and **J. J. Corso**. Actionness ranking with lattice conditional ordinal random fields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[Dav01] D. Davidson. Actions, reasons and causes (1963). In *Essays on Actions and Events*. Clarendon Press, Oxford, 2001.

[DBD+13] J. A. Delmerico, D. Baran, P. David, J. Ryde, and **J. J. Corso**. Ascending stairway modeling from dense depth imagery for traversability analysis. In *Proceedings of IEEE International Conference on Robotics and Automation*, 2013.

[D**Corso**B+12] J. A. Delmerico, **J. J. Corso**, D. Baran, P. David, and J. Ryde. Ascending stairway modeling: A first step toward autonomous multi-floor exploration. In *Proceedings of IEEE/RSJ Intelligent Robots and Systems (Video Proceedings)*, 2012.

[D**Corso**BD13] J. A. Delmerico, **J. J. Corso**, D. Baran, and P. David. Towards autonmous multi-floor exploration: Ascending stairway localization and modeling. Technical Report ARL-TR-6381, Army Research Laboratory, 2013.

[DD**Corso**11] J. A. Delmerico, P. David, and **J. J. Corso**. Building facade detection, segmentation, and parameter estimation for mobile robot localization and guidance. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2011.

[Del13] J. A. Delmerico. *Attributed Object Maps: Descriptive Object Models as High-level Semantic Features for Mobile Robotics*. PhD thesis, State University of New York at Buffalo, 2013.

[DKD**Corso**14] V. Dhiman, A. Kundu, F. Dellaert, and **J. J. Corso**. Modern MAP inference methods for accurate and faster occupancy grid mapping on higher order factor graphs. In *Proceedings of International Conference on Robotics and Automation*, 2014.

[DR**Corso**13] V. Dhiman, J. Ryde, and **J. J. Corso**. Mutual localization: Two camera relative 6-dof pose estimation from reciprocal fiducial observation. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2013.

[DRMS07] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

[DSCW10] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[DXD**Corso**13] P. Das, C. Xu, R. F. Doell, and **J. J. Corso**. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[Elf89] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[EVGW+10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[FBCM04] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2), 2004.

[FBM01] Charless Fowlkes, Serge Belongie, and Jitendra Malik. Efficient spatiotemporal grouping using the Nyström method. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 231–238, 2001.

[FGMR10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[FH04] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[FOS09] A. Franchi, G. Oriolo, and P. Stegagno. Mutual localization in a multi-robot system with anonymous relative position measures. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3974–3980. IEEE, 2009.

[GFM11] Ross B. Girshick, Pedro F. Felzenszwalb, and David Mcallester. Object detection with grammar models. In *In NIPS*, 2011.

[GKHE10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[GLSU13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *I. J. Robotic Res.*, 32(11):1231–1237, 2013.

[GVH03] Brian Gerkey, Richard T Vaughan, and Andrew Howard. The player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings of the 11th International Conference on Advanced Robotics*, 2003.

[GZS] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *in IEEE Intelligent Vehicles Symposium, 2011*, pages 963–968.

[HLON94] B.M. Haralick, C.N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision*, 13(3):331–356, 1994.

[HR03] Andrew Howard and Nicholas Roy. The robotics data set repository (radish), 2003.

[HYL+07] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

[KFL01] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

[KGGHW12] Orit Klliper-Gross, Yaron Gurovich, Tal Hassner, and Lior Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012.

[KJG+11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *Proceedings of IEEE International Conference on Computer Vision*, 2011.

[KKJ11] A. Kundu, K.M. Krishna, and C.V. Jawahar. Realtime multibody visual SLAM with a smoothly moving monocular camera. In *ICCV*, 2011.

[KMM+13] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2013.

[KMS08] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of British Machine Vision Conference*, 2008.

[KP09] Nikos Komodakis and Nikos Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *CVPR*, 2009.

[Lap05] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 2005.

[Liu02] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2002.

[LKG11] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, 2011.

[LSXFF10] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Proceedings of Advance in Neutral Information Processing*, 2010.

[LX**Corso**15] J. Lu, R. Xu, and **J. J. Corso**. Human action segmentation with hierarchical supervoxel consistency. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Mar12] A. Martinelli. Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination. *Robotics, IEEE Transactions on*, (99):1–17, 2012.

[MB13] R.S. Merali and T.D. Barfoot. Occpancy grid mapping with markov chain monte carlo gibbs sampling. In *IROS*, 2013.

[MDBB12] Daniel Meyer-Delius, Maximilian Beinhofer, and Wolfram Burgard. Occupancy grid models for robot mapping in changing environments. In *AAAI*, 2012.

[ME85] H.P. Moravec and A. Elfes. High resolution maps from wide angle sonar. In *ICRA*, volume 2, pages 116–121, 1985.

[MHK06] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104:90–126, 2006.

[Mor88] Hans P Moravec. Sensor fusion in certainty grids for mobile robots. *AI magazine*, 9(2):61,

1988.

[MPK09] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proceedings of IEEE International Conference on Computer Vision*, 2009.

[NDI+11] R.A. Newcombe, A.J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[NUS12] KS Nagla, Moin Uddin, and Dilbag Singh. Improved occupancy grid mapping in specular environment. *Robotics and Autonomous Systems*, 60(10):1245 – 1252, 2012.

[Par08] S. Paris. Edge-preserving smoothing and mean-shift segmentation of video streams. In *Proceedings of European Conference on Computer Vision*, 2008.

[PD07] S. Paris and F. Durand. A topological approach to hierarchical segmentation using mean shift. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[Pot07] Brian Potetz. Efficient belief propagation for vision using linear constraint nodes. In *CVPR*, 2007.

[QL99] L. Quan and Z. Lan. Linear n-point camera pose determination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(8):774–780, 1999.

[RAS08] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[R**Corso**12] J. Ryde and **J. J. Corso**. Fast voxel maps with counting bloom filters. In *Proceedings of International Conference on Intelligent Robots and Systems*, 2012.

[RF03] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *Proceedings of Advance in Neutral Information Processing*, 2003.

[RM03] X. Ren and J. Malik. Learning a classification model for segmentation. In *Proceedings of International Conference on Computer Vision*, volume 1, pages 10–17, 2003.

[RX02] H. Ren and G. Xu. Human action recognition in smart classrooms. In *Proceedings of IEEE International Conference on Face and Gesture*, 2002.

[SBB00] E. Sharon, A. Brandt, and R. Basri. Fast Multiscale Image Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 70–77, 2000.

[S**Corso**12] S. Sadanand and **J. J. Corso**. Action bank: A high-level representation of activity in video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[SGJ11] David Sontag, Amir Globerson, and Tommi Jaakkola. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1, 2011.

[SGS+06] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006.

[SLC04] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proceedings of International Conference on Pattern Recognition*, 2004.

[SM97] C. Schmid and R. Mohr. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[SM00] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[Sug88] Kokichi Sugihara. Some location problems for robot navigation using a single camera. *Computer Vision, Graphics, and Image Processing*, 42(1):112 – 129, 1988.

[Thr02] Sebastian Thrun. Robotic mapping: A survey. *Exploring Artificial Intelligence in the New Millenium*, 2002.

[Thr03] Sebastian Thrun. Learning occupancy grid maps with forward sensor models. *Autonomous Robots*, 15(2):111–127, 2003.

[TZZR10] N. Trawny, X.S. Zhou, K. Zhou, and S.I. Roumeliotis. Interrobot transformations in 3-D. *Robotics, IEEE Transactions on*, 26(2):226–243, 2010.

[UCF] `http://server.cs.ucf.edu/~vision/data.html`.

[WKSL11] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

[WUK+09] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of British Machine Vision Conference*, 2009.

[XAK+12] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and **J. J. Corso**. Combining skeletal pose with local motion for human activity recognition. In *Proceedings of VII Conference on Articulated Motion and Deformable Objects*, 2012.

[X**Corso**12a] C. Xiong and **J. J. Corso**. Coaction discovery: Segmentation of common actions across multiple videos. In *Proceedings of Multimedia Data Mining Workshop in Conjunction with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (MDMKDD)*, 2012.

[X**Corso**12b] C. Xu and **J. J. Corso**. Evaluation of super-voxel methods for early video processing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[XHX**Corso**15] C. Xu, S.-H. Hsieh, C. Xiong, and **J. J. Corso**. Can humans fly? Action understanding with multiple classes of actors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[XJ**Corso**12a] C. Xiong, D. Johnson, and **J. J. Corso**. Efficient max-margin metric learning. In *Proceedings of European Conference on Data Mining*, 2012. **Winner of Best Paper Award at ECDM 2012.**

[XJ**Corso**12b] C. Xiong, D. Johnson, and **J. J. Corso**. Spectral active clustering via purification of the k-nearest neighbor graph. In *Proceedings of European Conference on Data Mining*, 2012.

[XJ**Corso**14] C. Xiong, D. M. Johnson, and **J. J. Corso**. Active clustering with model-based uncertainty reduction. Technical Report 1402.1783, arXiv, 2014.

[XJX**Corso**12] C. Xiong, D. Johnson, R. Xu, and **J. J. Corso**. Random forests for metric learning with implicit pairwise position dependence. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[XX**Corso**12] C. Xu, C. Xiong, and **J. J. Corso**. Streaming hierarchical video segmentation. In *Proceedings of European Conference on Computer Vision*, 2012.

[ZR10] Xun S Zhou and Stergios I Roumeliotis. Determining the robot-to-robot 3D relative pose using combinations of range and bearing measurements: 14 minimal problems and closed-form solutions to three of them. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2983–2990. IEEE, 2010.

[ZR12] X. S. Zhou and S. I. Roumeliotis. Determining 3-D relative transformations for any combination of range and bearing measurements. *Robotics, IEEE Transactions on*, PP(99):1–17, 2012.

[ZT12] D. Zou and P. Tan. CoSLAM: Collaborative visual SLAM in dynamic environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.