Boosting with Stereo Features for Building Facade Detection on Mobile Platforms

Jeffrey A. Delmerico and Jason J. Corso
Department of Computer Science and Engineering
SUNY at Buffalo
Email: {jad12 , jcorso}@buffalo.edu

Philip David
Army Research Laboratory
Adelphi, Maryland
Email: phild@arl.army.mil

***Abstract*—Boosting has been widely used for discriminative modeling of objects in images. Conventionally, pixel- and patch-based features have been used, but recently, features defined on multilevel aggregate regions were incorporated into the boosting framework, and demonstrated significant improvement in object labeling tasks. In this paper, we further extend the boosting on multilevel aggregates method to incorporate features based on stereo images. Our underlying application is building facade detection on mobile stereo vision platforms. Example features we propose exploit the algebraic constraints of the planar building facades and depth gradient statistics. We've implemented the features and tested the framework on real stereo data.**

I. Introduction

Accurate scene labeling can enable applications which rely on the semantic information in an image to make high level decisions. Our goal of labeling building facades is motivated by the problem of mobile robot localization, which, in the framework currently being developed by our group, depends on detection of buildings within the field of view of the cameras on a mobile platform. Within this problem, accurate detection and labeling is critical for the high level localization tasks. Our approach uses multilevel 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. These aggregate features are in addition to pixel- and patch-level features, and are all used within an AdaBoost [1] two- or multi-class discriminative modeling framework. This recently-proposed method, Boosting on Multilevel Aggregates (BMA) [2], exploits the propensity of these coarsened regions to adhere to object boundaries, which in addition to the expanded feature set, offer less polluted statistics than patch-based features, which may violate those boundaries.

The Boosting on Multilevel Aggregates method provides two primary enhancements to the stock AdaBoost method. First, adaptive coarsening groups pixels into regions of similar intensity. Higher up the hierarchy, it groups these regions together, such that each aggregate represents a node in a graph with all of its associated pixels or aggregates from the next finer level as its *child* nodes. The multilevel aggregate hierarchy allows for statistics to be computed across regions which are less likely to violate object boundaries due to the adaptive nature of the coarsening procedure, and it enables the modeling framework to capture the properties of each class at various scales. Secondly, the aggregates offer a richer set of statistical features with which to model the classes, including shape, contextual, and hierarchical features. So in the weak learner set, in addition to position, patch-based histograms, and Haar-like and Gabor filters, we can include averages, moments, and histograms over the aggregates, as well as shape and context features for the aggregate regions such as rectangularity and number of neighbors. Training and classification proceed in the same manner as standard AdaBoost, except that the graph hierarchies are built on top of each image, and the expanded feature set is applied to the aggregates in the hierarchy when learning the model and when performing pixel labeling on new images.

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. A problem-specific feature of building facade detection is that, generally speaking, buildings tend to have planar surfaces on their exteriors. We can exploit this property of our building class as a feature in our AdaBoost framework by recognizing that planar surfaces can be represented as linear functions in disparity space and thus have constant spatial gradients [3]. Our primary contribution is the expansion of the BMA framework to include hierarchical disparity features in addition to BMA's appearance features. We apply all of the same pixel-, patch-, and aggregate-based features of BMA to the disparity map, as well as a set of new disparity-specific features intended to capture the planarity of building facades.

A. Related Work

Other research in the area of mobile robot localization from stereo cues includes the work of Konolige et al. [4], which integrates appearance and disparity information for object avoidance, and uses AdaBoost to learn color and geometry models for ideal routes of travel along the ground. They use stereo information for detection of the ground plane and for distinguishing obstacles, but not for classifying and labeling those objects. Luo and Maître [5] proposed using the same algebraic constraint on planar surfaces, but for the purpose of correcting disparity. Their approach relies on the assumption that within urban scenes, all surfaces will be planes, so their geometric properties can be used to enhance

978-1-4244-9300-5/10/$26.00 ©2010 IEEE

poor disparity calculations. Instead, we are using the linear gradient constraint on planar surfaces to identify those regions which do, in fact, fit that planar assumption. Li et al. [6] use disparity data in a template-based AdaBoost framework. Their work is applied to human pose estimation, and their features are strictly pixel-based. Perhaps the most similar approach to ours is from Walk et al. [7], which incorporates object-specific stereo features into a combination of classifiers for pedestrian detection. Although these disparity features are very different from the ones that we use, the use of object-specific properties to drive those features is consistent with our approach. However, their ultimate goal is for detection of pedestrian bounding boxes, and not for pixel labeling of those detected pedestrians. An important distinction between the two problems is also that buildings can occupy a much larger percentage of the pixels in the frame, and come in a much greater range of shapes, sizes, and appearances than humans.

II. Methods

We implement the standard Boosting on Multilevel Aggregates algorithm described in [2], but with extensions for working with disparity maps and their associated features. These additions include accommodations for working with invalid data in the disparity map: areas of the scene outside the useful range of the stereo camera, and dropouts where the disparity can not be computed within the camera's range due to occlusion or insufficient similarity between the images for a match at that point.

A. Dense Disparity

Computing the dense disparity map of a scene, given a stereo pair, is a non-trivial problem [8]. Many commercial stereo cameras are equipped with embedded processing for real-time disparity map computation. Although these products often have good resolution and do a decent job of computing the disparity map, there are limitations inherent in both the hardware and software. Stereo cameras generally have fixed focal length sensors, so the range in which the cameras can focus is limited, resulting in a finite region in which disparity can accurately be computed. Additionally, the on-board processors of stereo cameras can not execute the more accurate, but computationally intensive, disparity map algorithms such as TRW-S [9]. Even off-line computation of the disparity map is imperfect, because occluded regions from one image will not have a match in the other image, and thus will not have a disparity value. Figure 1 illustrates a typical example of a disparity map with invalid regions (shown in black). We discuss our accomodations for these obstacles in sections II-B and II-D.

B. Coarsening on Disparity

We perform coarsening on the disparity map in the same manner as the image intensity coarsening procedure proposed in [2]. Invalid disparities are first mapped to zero, and then a hierarchy of equal height to the image-based hierachy is built. We use the same definition of pixel affinity as [2] does for intensity: $exp[-|s_u - s_v|]$ for pixels/aggregates u and v, and their associated statistics s, which in this case is disparity. An example of intensity and disparity hierarchies produced by this procedure is illustrated in Figure 2. Although the coarsening proceeds similarly for both intensity and disparity, and the aggregates for both still tend to adhere to object boundaries, the resulting hierarchies have somewhat different character. The separate disparity hierarchy allows the aggregate features to capture the statistics of regions with similar disparity values, which may not align with regions of similar intensity.

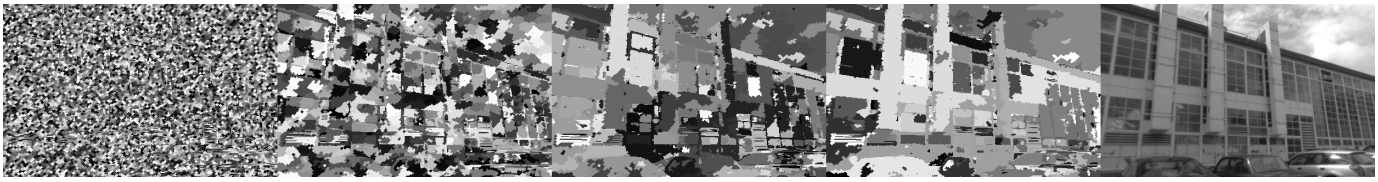
Fig. 1. A typical image with its disparity map. Invalid regions of the disparity map are in black.

C. Disparity Features

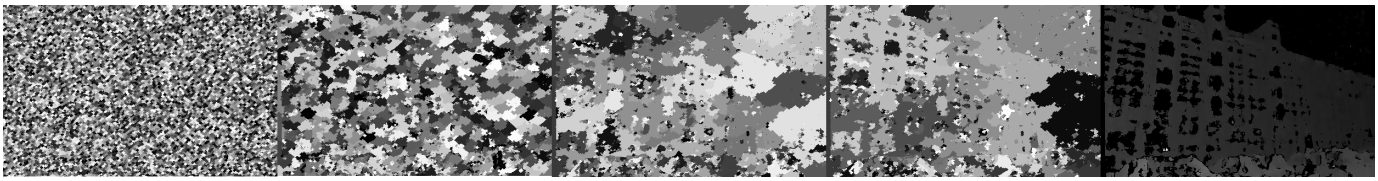
The BMA framework for intensity images adds a variety of aggregate features to the pixel- and patch-based statistics of standard boosting [2]. Aggregate spatial and photometric statistics include averages, moments, and histograms of intensities, colors, and Gabor responses. Shape features include elongation (ratio of bounding box height to width), rectangularity (amount of bounding box filled by aggregate), and several PCA statistics. Region and context features include adaptive Haar-like features on the aggregate bounding boxes at each level of the hierarchy, and several measures of neighbor similarity and region homogeneity. Other features such as aggregate mass and number of neighbors also help to capture some of the region-level information, in this case some of the hierarchical properties of the aggregates. We implement all of these pixel-, patch-, and aggregate-based features for disparity, in addition to several disparity-specific features intended to help discriminate between building and non-building pixels by measuring the uniformity of the disparity gradient across an aggregate [3]. We compute the x and y gradient images of the disparity by filtering with directional derivatives of gaussian kernels. From these gradient images, we compute the average and range of the gradient in each direction, as well as the vector gradient magnitude and angle. We have also included the Lapacian as a feature, because the Lapacian of a planar surface in a disparity map is zero.

D. Training and Classification

When we wish to classify an image, some regions will not have corresponding disparities; we compensate by basing our classification scheme on two models. We use a model that includes both image and disparity features for classifying pixels which do have valid disparity values, and a second model with only image features for classifying the pixels in invalid disparity regions. We train both models on pixels and

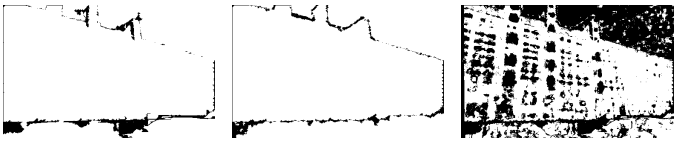


(a) Intensity hierarchy and original image (far right)



(b) Disparity hierarchy and disparity map (far right)

Fig. 2. Intensity and disparity hierarchies. Aggregate regions are colored with random gray values.



(a) Same label in image hierarchy.

(b) Same label in disparity hierarchy.

(c) Same label in both hierarchies, and valid disparity.

Fig. 3. Suitable pixels for training the image+disparity model (in white).

their corresponding aggregates from a single set of training images; in both cases, we only use a pixel if it has a consistent class label in all of the associated aggregates above it in the hierarchy. This avoids training on pixels whose aggregate statistics may be polluted at some higher level. For the image+disparity model, we further constrain the set of suitable training pixels by applying the same criteria to the labels up the disparity hierarchy, and by restricting the set to those pixels that have valid disparity values, as in Figure 3. Since we are using the image-only model to classify those pixels that do not have valid disparity, we train the image model on those pixels that have consistent labels in both hierarchies and invalid disparity in the training data. So during classification, given an input image and disparity map, pixels from valid regions of the disparity map are classified using the model incorporating both image and disparity features, and pixels in invalid regions are classified using the model with only image features.

III. Experimental Results

We have implemented and tested the described system on the two-class problem of building facade labeling. During training, we assembled a set of approximately 7500 weak classifiers, including pixel-, patch-, and aggregate-based features for both image and disparity statistics. During the boosting procedure, 200 of these weak classifiers are automatically selected for the discriminative model. For the image+disparity model, approximately 30% of the features selected during boosting were disparity features; about 17% were pixel- or patch-based features. The high rate of selection of disparity features, including several of the disparity-specific gradient features, indicates that these features provide additional discriminating power over the image features alone. The remainder of this section will be devoted to analyzing the performance of our proposed two-model approach relative to an image-only BMA classifier.

We used a data set captured by our team of a variety of natural scenes on a university campus. It features varied building types and styles, and the buildings are at different scales and orientations. Our dataset consists of 142 grayscale stereo pairs with associated 16-bit disparity maps, which we split into a randomly selected training set of 100 images, and a validation set of the remaining 42 images. We trained the two models for the image+disparity system on the training data, as well as a standard appearance-based BMA model for comparison.

A comparison of overall accuracy for standard BMA and our Image+Disparity BMA model is presented in Table I. Although the overall gains in accuracy are modest ($\sim$2%), we do demonstrate a significant ($\sim$8%) increase in the accuracy of labeling the positive case (building). The low overall improvement, however, is due to some loss in accuracy with the negative, or background (BG) case. However, given our intended application, a higher rate of false positives is tolerable, because noisy output can be improved in post-processing. Some of these shortcomings are likely due to the quality and quantity of the data set. We intend to expand our data set in the future with more captured images, and we hope to improve our existing data with more accurate offline disparity map computation. We are also currently pursuing some post-processing steps to use inference to improve the results, as well as to enforce the planar surface constraint on the labeled regions. Figure 4 presents some examples of the output for both the standard BMA model and our image+disparity BMA model. For several of the images, there is noticeable improvement in the labeling accuracy (namely the last two), but for the others, there are improvements in some areas with losses in others. In many instances, we also note increased confidence in labeling both classes, as shown in the probability images. Part of our desire to develop a larger, more varied data set is to expand the variety of building styles and textures available to us for

Fig. 4. Several examples of classification output. For each image set, the individual images are (L to R): Original Image, Standard BMA Probability, Image+Disparity BMA Probability, Ground Truth, Standard BMA Label, Image+Disparity BMA Label

TABLE I
CONFUSION MATRICES FOR THE STANDARD BMA MODEL, AND OUR IMAGE+DISPARITY BMA MODEL.

	Standard BMA		Image+Disparity BMA	
	BG	Building	BG	Building
BG	77.12%	22.89%	74.28%	25.72%
Building	26.68%	73.32%	18.94%	81.05%
Total Accuracy	75.43%		77.27%	

training the model, as well as to incorporate more negative (background) data, in order to cope with some of the issues that appear to cause difficulty for our models (i.e. the rock and water tank in the middle image).

IV. CONCLUSION

We present a extension of the Boosting on Multilevel Aggregates method [2] to include features on disparity map data from stereo cameras, in the context of building facade detection for mobile robot platforms. Our approach achieves significant improvement in the accuracy of labeling the positive (building) class, albeit with some loss of accuracy in labeling the negative class (background), compared to the BMA method using appearance-based features alone.

We intend to pursue several paths toward improving our results: expansion and enhancement of our data set and post-processing steps to refine the output. Ultimately, we plan to embed this system on an actual mobile platform in order to incorporate our building facade detection into a mobile robot localization system.

ACKNOWLEDGMENT

We are grateful for the financial support provided in part by NSF CAREER IIS-0845282 and DARPA/ARL Mind's Eye W911NF-10-2-0062. A portion of this work was completed while J.D. was an intern at the United States Army Research Laboratory.

REFERENCES

[1] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[2] J. Corso, "Discriminative modeling by boosting on multilevel aggregates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] J. Corso, D. Burschka, and G. Hager, "Direct plane tracking in stereo images for mobile navigation," in *IEEE International Conference on Robotics and Automation*, 2003.

[4] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, and B. Gerkey, "Outdoor mapping and navigation using stereo vision," in *Proceedings of the Intl. Symp. on Experimental Robotics*, 2008.

[5] W. Luo and H. Maitre, "Using surface model to correct and fit disparity data in stereo vision," in *Proceedings of the 10th International Conference on Pattern Recognition*, 1990.

[6] L. Li, K. Hoe, X. Yu, L. Dong, and X. Chu, "Human Upper Body Pose Recognition Using Adaboost Template For Natural Human Robot Interaction," in *Proceedings of Canadian Conference on Computer and Robot Vision*, 2010.

[7] S. Walk, K. Schindler, and B. Schiele, "Disparity Statistics for Pedestrian Detection: Combining Appearance, Motion and Stereo," in *Proceedings of European Conference on Computer Vision*, 2010.

[8] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[9] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1568–1583, 2006.