Building Facade Detection, Segmentation, and Parameter Estimation for Mobile Robot Stereo Vision

Jeffrey A. Delmerico[1a], Philip David[b], Jason J. Corso[a]

[a] SUNY Buffalo, Department of Computer Science and Engineering, 338 Davis Hall, Buffalo, NY, 14260-2000, {jad12,jcorso}@buffalo.edu
[b] Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783-1197, philip.j.david4.civ@mail.mil

Abstract

Building facade detection is an important problem in computer vision, with applications in mobile robotics and semantic scene understanding. In particular, mobile platform localization and guidance in urban environments can be enabled with accurate models of the various building facades in a scene. Toward that end, we present a system for detection, segmentation, and parameter estimation of building facades in stereo imagery. The proposed method incorporates multilevel appearance and disparity features in a binary discriminative model, and generates a set of candidate planes by sampling and clustering points from the image with Random Sample Consensus (RANSAC), using local normal estimates derived from Principal Component Analysis (PCA) to inform the planar models. These two models are incorporated into a two-layer Markov Random Field (MRF): an appearance- and disparity-based discriminative classifier at the mid-level, and a geometric model to segment the building pixels into facades at the high-level. By using object-specific stereo features, our discriminative classifier is able to achieve substantially higher accuracy than standard boosting or modeling with only appearance-based features. Furthermore, the results of our MRF classification indicate a strong improvement in accuracy for the binary building detection problem and the labeled planar surface models provide a good approximation to the ground truth planes.

Keywords: stereo vision, mobile robot perception, hierarchical Markov random field, building facade detection, model-based stereo vision

1. Introduction

Accurate scene labeling can enable applications that 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 in GPS-denied areas, which commonly arises in urban environments. Besides GPS, other cues from the environment such as compass headings and Time-Difference-Of-Arrival (TDOA) of radio signals, along with vision-based localization [1], can enable semantic methods of navigation in these areas. However, these methods suffer from low accuracy and are subject to interference, or in the case of vision-based localization, struggle with occlusion and clutter in the scene. The vision-based localization approach being developed by our group depends on the detection of buildings within the field of view of the cameras on a mobile platform as a means to reduce the effects of clutter on localization, and to enable navigation based on static, semantically meaningful landmarks detected in the scene. Within this problem, accurate detection and labeling of the facades is important for the high level localization and guidance tasks. We restrict our approach to identifying only planar building facades, and we require image input from a stereo source that produces a disparity map. Since most buildings have planar facades, and many mobile robotic platforms are equipped with stereo cameras, neither of these assumptions is particularly restrictive.

In this paper, we propose a method for fully automatic building facade imaging–detection, segmentation, and parameter estimation–for mobile stereo vision platforms. For an input stereo image and disparity map, we desire a pixelwise segmentation of the major building facades in the scene, as well geometric models for each of these planar facades. Our approach proceeds in three main steps: discriminative modeling with both appearance and disparity features, candidate plane detection through PCA and RANSAC, and energy minimization of MRF potentials. A diagram of the workflow for candidate plane detection and high-level labeling is provided in Fig. 1. We make no assumptions on the quality of the input data, and in fact many of our methods were driven by the need to deal with the missing or inaccurate data that is common to single-view stereo imagery. Consequently, we adopt a top-down approach to fitting planes globally in the image, rather than a bottom-up approach that would suffer from missing disparity data on the local scale. This is also directed toward our goal of segmenting the major facades in the scene, and not every planar surface. In our experiments, we use off-the-shelf single-view stereo data produced by a system-on-a-chip camera that computes disparity in real time, and we acknowledge that the maps may suffer from both missing data and range-uncertainty.

Our work leverages stereo information from the beginning.

[1] Present address: *University of Hawai'i at Manoa, Department of Mechanical Engineering, 2540 Dole St.-Holmes Hall 310, Honolulu, HI 96822,* jad4@hawaii.edu

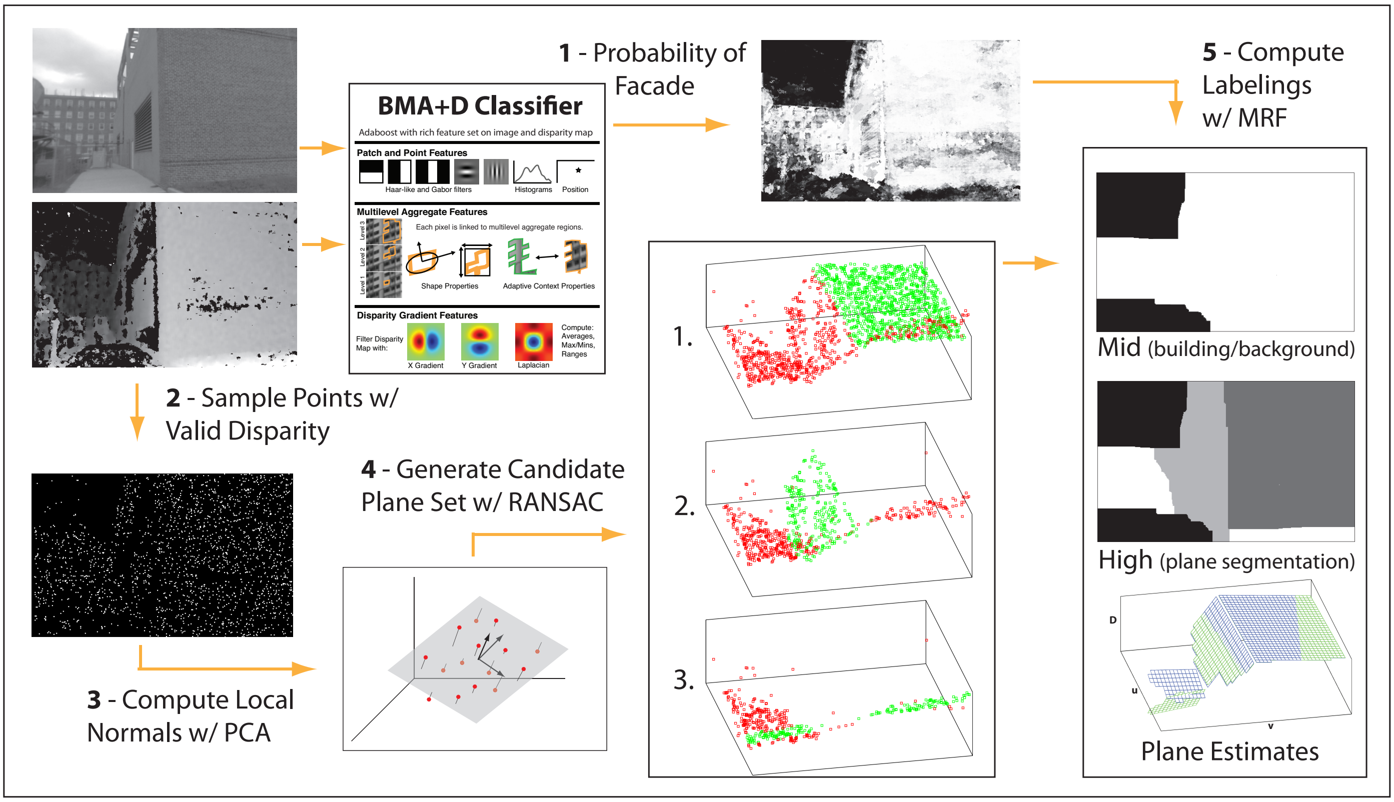

1 - Probability of Facade
BMA+D Classifier
Adaboost with rich feature set on image and disparity map
Patch and Point Features
Haar-like and Gabor filters
Histograms
Position
Multilevel Aggregate Features
Level 3
Level 2
Level 1
Each pixel is linked to multilevel aggregate regions.
Shape Properties
Adaptive Context Properties
Disparity Gradient Features
Filter Disparity Map with:
X Gradient
Y Gradient
Laplacian
Compute: Averages, Max/Mins, Ranges
5 - Compute Labelings w/ MRF
2 - Sample Points w/ Valid Disparity
4 - Generate Candidate Plane Set w/ RANSAC
1.
2.
3.
Mid (building/background)
High (plane segmentation)
D
u
v
Plane Estimates
3 - Compute Local Normals w/ PCA

Figure 1: Workflow of the proposed method. The proposed BMA+D classifier computes a probability map for binary classification of pixels into buildings and non-buildings (Step 1, Sec. 3). We then generate a set of candidate planes with parameter estimates using a RANSAC model that incorporates local PCA normal approximations (Steps 2-4, Sec. 4.2). Finally, we solve a two-layer MRF to compute labelings for the binary classification at the mid-level and for facade segmentation at the high-level (Step 5, Sec. 4.3).

Our discriminative model is generated from an extension of the Boosting on Multilevel Aggregates (BMA) method [2] that includes stereo features [3]. 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 [4] two- or multi-class discriminative modeling framework. 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, we use the stereo features to exploit the property that planes can be represented as linear functions in disparity space and thus have constant spatial gradients [5]. We will refer to this extension of BMA to disparity features as BMA+D. We use the discriminative classification probability as a prior when performing inference for the facade labels.

In order to associate each building pixel with a particular facade, we must have a set of candidate planes from which to infer the best fit. We generate these planes by sampling the image and performing Principal Component Analysis (PCA) on each local neighborhood to approximate the local surface normal at the sampled points. We then fit models to those points by iteratively using Random Sample Consensus (RANSAC) [6] to find subsets that fit the same plane and have similar local normal orientations. From these sets of points, we are able to estimate the parameters of the primary planes in the image.

We then incorporate both of these sources of information into a Bayesian inference framework using a two-layer Markov Random Field (MRF). We represent the mid-level MRF as an Ising model, a layer of binary hidden variables representing the answer to the question “Is this pixel part of a building facade?” This layer uses the discriminative classification probability as a prior, and effectively smooths the discriminative classification into coherent regions. 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. For each pixel, we consider its image coordinates and disparity value, and evaluate the fitness of each candidate plane to that pixel, and incorporate it into the energy of labeling that pixel as a part of that plane. A more in-depth discussion of our modeling and labeling methods can be found in Section 4.

The primary contributions of this paper 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 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. Our method for facade segmentation using the two-layer MRF and RANSAC was originally proposed in [7], but this paper includes a full quantitative study on the performance of these methods on a larger dataset, and this is the first inclusion of any of this work in an archival publication.

1.1. Related Work

Other research in the area of modeling with stereo cues includes the work of Konolige et al. [8], 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. Li et al. [9] 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 our discriminative modeling method is from Walk et al. [10], 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 and sizes than humans.

Luo and Maître [11] 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 poor disparity calculations. Instead, we are using a linear gradient constraint in our disparity features to identify those regions which do, in fact, fit that planar assumption.

Building facade detection and segmentation have been and continue to be well-studied problems. Many recent papers in the literature have focused on segmentation of building facades for use in 3D model reconstruction, especially in the context of architectural modeling or geo-spatial mapping applications such as Google Earth. Korah and Rasmussen use texture to segment building facades, among other facade-related tasks [12]. Frohlich et al. [13] perform facade labeling with a Randomized Decision Forest, but do not attempt to segment individual facades. Wendel at al. [14] use intensity profiles to find repetitive structures in coherent regions of the image in order to segment and separate different facades. Hernández and Marcotegui employ horizontal and vertical color gradients, again leveraging repetitive structures, to segment individual facades from blocks of contiguous buildings in an urban environment. Hoeim et al. [15] use a single camera image to infer coarse planar orientations for regions of the image in order to create popped-up 3D views of the scene, but their approach does not consider segmentation or modeling of buildings or their facades. Recky et al. [16] use semantic segmentation of the scene with a discriminative random field, then find repetitive patterns and leverage some contextual constraints (e.g. facade boundaries will be vertical) to compute facade-wise segmentation. However, their impressive results (96.6% F-score) require multi-view. With single-view, their approach produces comparable results to ours (81.7% pixel-wise F-score vs. our 77.7%). Although they are interested in facade segmentation of the images, they do not pursue any disparity or depth information from their multi-view scenario, and thus do not attempt any modeling of the facades that they segment. The multiview approach in [17] automatically creates textured 3D models of urban scenes from sequences of images. They perform semantic segmentation on the images and partition the resulting 3D facades along vertical lines between buildings. They produce a very realistic looking 3D model for each building by leveraging the regularity of buildings in urban areas. However, there are no quantitative results with which to compare our performance.

Despite the additional information that multi-view stereo provides, we pursue a single-view approach due to our problem constraints. For image-based localization from facade estimates, we anticipate the need to capture many single stereo frames in a panorama. Facade orientations within the narrow field of view of a single stereo image likely will not constrain the location or pose of the camera with respect to the buildings in an urban environment. However, by foveating to observe other buildings in a panorama, a set of facade estimates from multiple single-view stereo images can be pieced together to give a more constraining set of facades from a wider field of view. Additionally, many semantic scene segmentation approaches exist using single-view camera imagery. By utilizing depth from stereo, those single-view approaches can be extended to extract geometric information about the labeled facades in the form of planar models.

The homography approach as in [18] could be applied to this problem in order to bypass the disparity map altogether to obtain planar correspondences between images. However, we are pursuing a purely automatic approach that does not use prior knowledge or human intervention, and their real quadratic embedding approach requires the number of planes to be known a priori, and their feature points are hand-selected.

The approach in [19] uses appearance, stereo, and 3D geometric features from a moving camera with structure from motion. They leverage a Manhattan-world assumption in indoor scenes to achieve a three-class segmentation of the scene with $\sim 75\%$ labeling accuracy. Although their features and approach are very different from ours, and their problem more constrained, their use of stereo and 3D features in addition to visual features is in line with our proposed method.

Posner et al. [20] classify laser scan points that have been projected into the camera frame into 10 urban classes (e.g. brick, vehicle, grass). They take a bottom-up approach for plane fitting to their point cloud data: the space of the scan is discretized into cubes, and local plane models are fit to the points within them, and these local planes are merged into planar patches based on orientation. The plane orientation relative to the ground becomes a feature, along with numerous color and texture features, for a multiclass SVM classifier. They achieve high accuracy (83 – 91% for different types of walls) in classifying the pixels corresponding to points from their laser scans, but they do not do produce a full segmentation of the image, or isolate individual facades.

Several other methods utilize vanishing points for planar surface detection. David identifies vanishing points in a monocular image by grouping line segments with RANSAC and then determines plane support points by the intersection of the segments that point toward orthogonal vanishing points, ultimately clustering them to extract the planes of the facade [21]. Bauer et al. [22] implement a system for building facade detection using vanishing point analysis in conjunction with 3D point clouds obtained by corresponding a sweep of images with known orientations. Lee et al. [23] use a line clustering-based approach, which incorporates aerial imagery, vanishing points, and other projective geometry cues to extract building facade textures from ground-level images, again toward 3D architectural models reconstruction.

Our work draws on the contributions of Wang et al. [24], whose facade detection method using PCA and RANSAC with LiDAR data inspired our approach with stereo images. Perhaps the approach most similar in spirit to ours is that of Gallup et al. [25], who also use an iterative method for generating candidate plane models using RANSAC, and also solve the labeling problem using graph cuts [26]. However, their approach relies on multiview stereo data and leverages photoconsistency constraints in their MRF model, whereas we perform segmentation with only single stereo images. In addition, on a fundamental level their method involves finding many planes that fit locally, and stitching them together in a bottom-up manner, whereas we aim to extract our planar models from the global data set, without an explicit restriction on locality. We present quantitative results on the accuracy of our planar modeling as well.

Although many of these results are directed toward 3D model reconstruction, some other work has been focused toward our intended application of vision-based navigation, namely [21, 27, 28, 29]. Additionally, our work is focused on retrieval of the estimated plane parameters, as implemented in the planar surface model of [5], and not on 3D model reconstruction.

Our approach proceeds in five steps: 1) computing a probability map with a discriminative classifier (Sec. 3), 2) sampling the disparity map (Sec. 4.2.1), 3) computing local normal estimates at the sampled points using PCA (Sec. 4.2.1), 4) iteratively generating a set of candidate planes with RANSAC (Sec. 4.2.2), and 5) using a hierarchical Markov random field to compute facade segmentations (Sec. 4.3). Please see Fig. 1 for a visual representation of this workflow.

2. Boosting on Multilevel Aggregates

Our discriminative modeling approach is based on the Boosting on Multilevel Aggregates (BMA) method proposed in [2]. We use the version of BMA that is extended to include disparity features (BMA+D, see Sec. 3) for producing pixelwise probabilities for the building class. Although the full methodology is not reproduced here, the core components upon which our contributions are based are described below. The central idea of BMA is that the feature types that are traditionally used within a boosting framework (point and patch-based features), are limited in their discriminative power. Since these features do not leverage any context from the underlying image, their statistics are often polluted when the patches capture regions that contain pixels from multiple classes. In order to provide features that avoid this problem, and that also offer a richer set of statistics to measure from the image, BMA uses adaptive coarsening to build a hierarchy of aggregate regions from the image, essentially a hierarchy of linked superpixels. It links each pixel with the aggregates above it in the hierarchy, and computes features on the aggregates as well as the traditional patch and point-based features on the image. These aggregate features are rich in information that is not captured in the image-level features, they are computed at multiple scales, and they adapt to the underlying structure of the image to follow object boundaries. All of the new aggregate features, as well as patch-based Haar features and x and y coordinate point features, are used to train an AdaBoost model for discriminative classification.

2.1. Adaptive Multilevel Coarsening

From a graph defined on the image pixels, we compute a hierarchy of progressively coarser graph layers containing aggregate nodes grouped from the nodes of the finer layer. At each iteration of coarsening, each node in the current layer, representing a pixel or aggregate, is grouped with its connected neighbors into an aggregate based on the affinities of their statistics (e.g. intensity). Each aggregate inherits both connectivity and statistics from its children, the latter being the weighted mean of its children's properties, and all of its features are computed during coarsening. A reduction factor, τ, limits the number of children per aggregate, and therefore determines the height of the hierarchy. Coarsening is stable: the grouping procedure is based on a deterministic decision boundary for aggregate statistical affinity. In the worst case, the complexity of the coarsening procedure is log-linear ($O(n \log_{\frac{1}{\tau}} n)$) in the number of pixels, n, but linear ($O(n)$) in the average case. This coarsening procedure and the aggregate features summarized below are described in full detail in [2].

2.2. Aggregate Features

The features below are defined on the aggregates at each level of the hierarchy for an aggregate u using the following notation:

$L(u)$	set of pixels it represents
$N(u)$	set of neighbors on same level
$C(u)$	set of child nodes
$min_x(u), min_y(u)$	minimum spatial location
$max_x(u), max_y(u)$	maximum spatial location
$x(u), y(u)$	spatial location
$g(u), a(u), b(u)$	intensity and color (Lab space)

Photometric and Spatial Statistical Features

- **Average Statistics:** weighted mean for $x(u)$ (similarly for y, g, a, and b)

$$m(u) = \sum_{c \in C(u)} m(c) \tag{1}$$

$$x(u) = \frac{1}{m(u)} \sum_{c \in C(u)} m(c)x(c) \qquad (2)$$

- **Aggregate Moments:** central moment about the aggregate's mean statistic, computed over its set of pixels, for $x(u)$ (and similarly for y, g, a, and b)

$$M_x^k(u) = \frac{1}{m(u)} \sum_{i \in L(u)} (x(i) - x(u))^k \qquad (3)$$

- **Adaptive Histograms:** for intensity, colors, and Gabor responses are computed over $L(u)$. Histogram bin weights are each considered features. For example, bin b of the intensity histogram H_g:

$$H_g(u, b) = \frac{1}{m(u)} \sum_{i \in L(u)} \delta(g(i) - b) \qquad (4)$$

Shape Features

- **Elongation:** ratio of height to width of an aggregate's bounding box

$$e(u) = \frac{h(u)}{w(u)} = \frac{max_y(u) - min_y(u)}{max_x(u) - min_x(u)} \qquad (5)$$

- **Rectangularity:** measures the degree to which an aggregate fills its bounding box

$$r(u) = w(u)h(u) - m(u) \qquad (6)$$

- **PCA:** compute the 2D spatial covariance matrix and its two eigenvalues: $\lambda_1(u)$ and $\lambda_2(u)$. PCA features are $\lambda_1(u)$, $\lambda_2(u)$, the ratio $\frac{\lambda_2(u)}{\lambda_1(u)}$, and the off-diagonal covariance.

Adaptive Region and Contextual Features

- **Adaptive Relative Haar-like Features:** patch-based Haar features but with spatial coordinates defined relative to an aggregate's bounding box.

- **Contextual Features:** measure the similarity of an aggregate to its neighbors at a region level. Consider a distance measure $D(u, v)$ between neighboring aggregates u and v on a statistic (intensity for example). Define a min-context feature (and max and mean features similarly) as:

$$f(u) = \min_{v \in N(u)} D(u, v) \qquad (7)$$

Hierarchical Features

- **Mass:** $m(u)$ of an aggregate measures the homogeneity of a region.

- **Number of Neighbors:** $\mid N(u) \mid$ captures the local complexity of a region.

3. BMA+D Classifier

We implement the Boosting on Multilevel Aggregates algorithm described above, but with extensions for working with disparity maps and their associated features. This extension was initially proposed in [3] and expanded in [7]. In our facade segmentation algorithm, the BMA+D classifier produces a probability that each pixel in an input image is from the building class. The BMA method builds a hierarchy of aggregate regions on the input image and then uses novel features computed on these aggregate superpixel regions, in addition to pixel and patch based features, to perform discriminative classification within an AdaBoost framework. Our additions to BMA 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, lack of texture, or insufficient similarity between the images for a match at that point. Additionally, we introduce several novel disparity-based features into the boosting framework. The AdaBoost algorithm automatically selects the most discriminating features in an adaptive way, and produces the best subset of the full feature set, given the training data.

Although in principle any classifier could be used for this step, so long as it could produce a probability map for binary classification in identifying building pixels, we developed the BMA+Disparity Classifier as a way to incorporate problem-specific knowledge into the boosting framework. Our results in this domain are superior to approaches that do not leverage disparity information in their classification.

3.1. Dense Disparity

Computing the dense disparity map of a scene, given a stereo pair, is a non-trivial problem [30]. Although there have been recent advancements in sensors such as the Microsoft Kinect that produce very dense depth or disparity maps, and therefore enable high-level tasks that depend on the quality of that data (for example, [31]), these sensors are unsuitable for outdoor use. 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 [32]. 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 2 illustrates a typical example of a disparity map with invalid regions (shown in black). We discuss our accommodations for these obstacles in sections 3.2 and 3.4.

3.2. Coarsening on Disparity

We perform coarsening on the disparity map in the same manner as the image intensity coarsening procedure proposed

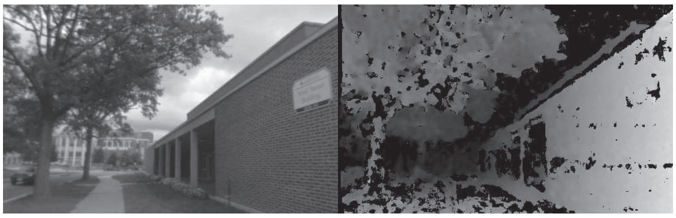


Figure 2: A typical image with its disparity map. Invalid regions of the disparity map are in black.

in [2]. Invalid disparities are first mapped to zero, and we then build a hierarchy of disparity aggregates of equal height to the one for the image. 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 3. 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.

3.3. Disparity Features

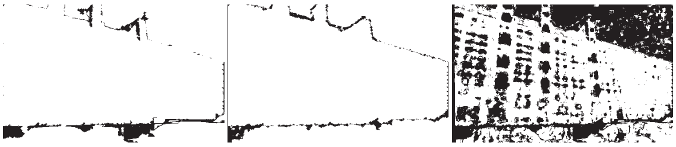
The BMA framework for intensity and color images adds a variety of aggregate features to the pixel- and patch-based statistics of standard boosting [2], all of which are summarized in Sec. 2 . We implement all of these pixel-, patch-, and aggregate-based features for disparity, and in addition include 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, we can separate the building and background classes by the property that planar facades will have constant gradient [5] in disparity space. We compute the x gradient images of the disparity map by filtering with the directional derivative of a 1-D Gaussian distribution in the x-direction (similarly for y):

$$\frac{\partial}{\partial x} G_{\sigma_x}(x) = -\frac{1}{\sqrt{2\pi}\sigma_x^3} x \exp\left(\frac{-x^2}{2\sigma_x^2}\right) \tag{8}$$

that is discretized into a kernel of fixed width. 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 Laplacian as a feature, because the Laplacian of a planar surface in a disparity map is zero. For this we convolve the image with the 3 × 3 Laplacian kernel.

3.4. 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 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 BMA+D 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 4. 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. As we are performing the coarsening procedure from standard BMA twice (once for the image and once for the disparity map), the complexity of this step is also log-linear ($O(n \log_{\frac{1}{\tau}} n)$) in the number of pixels n, in the worst case, and linear ($O(n)$) in the average case.



(a) Same label in image hierarchy.

(b) Same label in disparity hierarchy.

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

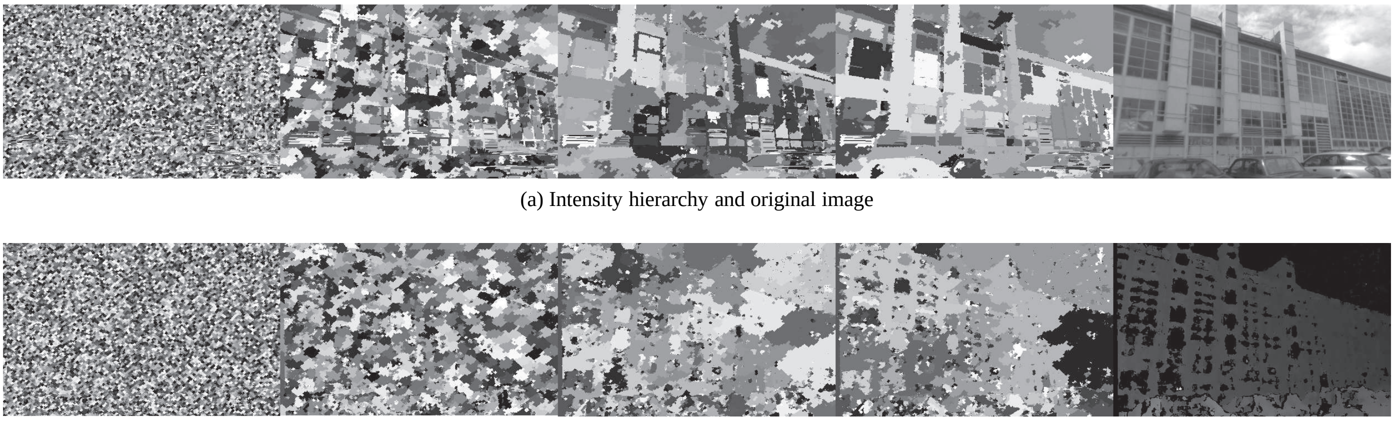
Figure 4: Suitable pixels for training the BMA+D model (in white).

4. MRF Model and Facade Parameter Estimation

We have developed a Markov random field model to perform segmentation and facade model labeling. For each pixel in an image, we compute its label for both the binary building/background labeling problem, as well as the best fit plane label among a set of facade models generated from the data. This overall approach was proposed initially in [7] but has been expanded and more thoroughly evaluated here.

4.1. Plane Parameters

We now derive the planar model that we use for modeling facades in disparity space. Throughout this discussion, we assume that we have stereo images in which the extrinsic calibration parameters are unknown but constant. Since we do not aim for full 3D reconstruction, we assume that the intrinsic calibration parameters are known to the camera or disparity source, but they are not required for modeling planes in disparity space given a disparity map. Thus, we can determine the surface normal parameters up to a constant that describes the camera parameters; and since that constant will be the same across all



(a) Intensity hierarchy and original image

(b) Disparity hierarchy and disparity map

Figure 3: Intensity and disparity hierarchies. The first four images in each row show the hierarchy levels from the lowest on the left (aggregates coarsened directly from source) to the highest on the right. The final image in each row is the source image in order to facilitate the comparison of object boundaries with top-level aggregate regions. At each level, the aggregate regions are colored with random gray values.

candidate planes, we can use the computed surface normals to differentiate between planes.

A plane in 3D space can be represented by the equation:

$$ax + by + cz = d \tag{9}$$

and for non-zero depth, z, this can be rewritten as:

$$a\frac{x}{z} + b\frac{y}{z} + c = \frac{d}{z} \tag{10}$$

We can map this expression to image coordinates by the identities $u = f \cdot \frac{x}{z}$ and $v = f \cdot \frac{y}{z}$, where f is the focal length of the camera. We can also incorporate the relationship of the stereo disparity value at camera coordinate (u, v) to the depth, z, using the identity $D(u, v) = \frac{fB}{z}$, where D is the disparity and B is the baseline of the stereo camera. Our plane equation becomes:

$$a\frac{u}{f} + b\frac{v}{f} + c = \frac{d \cdot D(u, v)}{fB} \tag{11}$$

which reduces to:

$$\left(\frac{aB}{d}\right)u + \left(\frac{bB}{d}\right)v + \left(\frac{cfB}{d}\right) = D(u, v) \tag{12}$$

Although $\mathbf{n} = (a, b, c)^T$ is the surface normal in world coordinates, for our purposes we can seek to determine the following uncalibrated plane parameters $\mathbf{n}' = (a', b', c')$, where:

$$a' = \frac{aB}{d}, b' = \frac{bB}{d}, c' = \frac{cfB}{d} \tag{13}$$

such that

$$\mathbf{n}' \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = a'u + b'v + c' = D(u, v) \tag{14}$$

This new set of plane parameters relates the image coordinates and their corresponding disparity values by incorporating the constant but unknown camera parameters.

4.2. *Candidate Plane Detection*

Our MRF computes the optimal label for each building pixel from a set of candidate planar models. We now describe the top-down approach that we use to generate the dominant planar models in an image.

We perform the second phase of our method by iteratively using RANSAC to extract a set of points that both fit a planar model in disparity space and have a local normal estimate that is consistent with the model. The extracted plane models become the set of candidate planes for our high-level MRF labeling. Each pixel in the image will be labeled by the MRF as belonging to one of these candidate planes or else assigned a null label.

4.2.1. *Local Normal Estimation*

Based on our assumption of planar building facades, we can use Principal Component Analysis to determine a local normal to a point in disparity space as in [33]. Since we are working with regionally dense disparity data, we sample from the available points that have valid disparity. For each sampled point, we first construct the covariance matrix for points in its neighborhood of the disparity map. To do this, we consider all points $p_i = (u_i, v_i, -D(u_i, v_i))$ with valid disparity in a 5 × 5 window centered on this point. Note that stereo cameras that compute the disparity map with onboard processing in real-time often do not produce dense disparity maps, so the neighborhood may be sparse. Other neighborhood sizes could be used, but we found that a 5 × 5 window provided good estimates while remaining local. We compute the centroid, $\overline{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$, of the points $\{p_i\}_{i=1\ldots N}$ in the neighborhood, and calculate the 3 × 3 covariance matrix with:

$$W = \frac{1}{N}\sum_{i=1}^{N}(p_i - \overline{p}) \otimes (p_i - \overline{p}) \tag{15}$$

where $\otimes$ is the outer product. We then compute the eigenvalues of W, and the eigenvectors corresponding to the largest two eigenvalues indicate the directions of the primary axes of a local planar estimate to that neighborhood of points. The eigenvector corresponding to the smallest eigenvalue thus indicates the direction of the local surface normal, $\mathbf{n}_{(u,v)}$.

4.2.2. RANSAC Plane Fitting

Once we have normal estimates, we take a greedy approach to fitting planar models to the points in disparity space, producing a set of models for the major planes in the image. We take a sample, S, of image points with valid disparity, and compute the local planar surface normal estimates by the aforementioned method. We then seek to fit a model to some subset of S of the form:

$$\alpha v + \beta u + \epsilon(-D(u, v)) + \theta = 0 \qquad (16)$$

where $\tilde{\mathbf{n}} = \frac{1}{\epsilon}(\alpha, \beta, \theta)$ is the surface normal from Eq. (14). Since RANSAC finds the largest inlier set, P_{in}, that it can among S, we will fit the most well-supported plane first [6]. We then remove the inliers, leaving $S' = S \setminus P_{in}$, and repeat this process iteratively, finding progressively less well-supported planes, until a fixed percentage of the original S has been clustered into one of the extracted planes. In our experiments, we used a sample of 2000 points from the image, and concluded the plane extraction once 80% of the points had been clustered, or when RANSAC failed to find a consensus set among the remaining points. We assume Gaussian noise on the inlier set for our RANSAC plane model, and throughout we use a standard deviation of $\sigma_\eta = 5$.

Although we use RANSAC to fit a standard plane model, we use a modified error term in order to incorporate the information in the local normal estimates. Here, since our local normal estimate required the use of a three dimensional coordinate system $(u, v, -D(u, v))$, and produces a normal of that form, we must use a slightly different normal formulation of $\mathbf{n}_m = (\alpha, \beta, \epsilon)$. The standard measure of error for a plane model is the distance of a point from the plane: $E_m = \mid \alpha v + \beta u + \epsilon(-D(u, v)) + \theta \mid$, assuming $\mathbf{n}_m = (\alpha, \beta, \epsilon)$ is a unit vector. We compute another measure of error, E_{norm}, the dot product of the model plane normal $\mathbf{n}_m$ and the local normal estimate $\mathbf{n}_{(u,v)}$, which is the cosine of the dihedral angle between the two planes defined by those normals. If we take its magnitude, this metric varies from 0 to 1, with 1 representing normals that are perfectly aligned, and 0 representing a dihedral angle of 90°. Since the range of E depends on the properties of the image (resolution, disparity range), we combine these two metrics as follows:

$$E = (2 - E_{norm})E_m = (2 - \mid \langle \mathbf{n}_m, \mathbf{n}_{(u,v)} \rangle \mid)E_m \qquad (17)$$

such that the dihedral angle scales the error term from E_m to $2E_m$, depending on the consistency of the model and local normals.

4.3. MRF Model

We model our labeling problem in an energy minimization framework as a pair of coupled Markov Random Fields. Our mid-level representation seeks to infer the correct configuration of labels for the question “Is this pixel part of a building facade?” Based on this labeling, the high-level representation seeks to associate those pixels that have been positively assigned as building facade pixels to one of the candidate planes identified by the previous RANSAC procedure. Figure 5 shows a graphical representation of this MRF model. Our motivation for this design stems from the fact that these are related but distinct questions, and they are informed by different approaches to modeling buildings. The mid-level MRF represents an appearance-based model, while the high-level MRF represents a generative model for the planar facades.

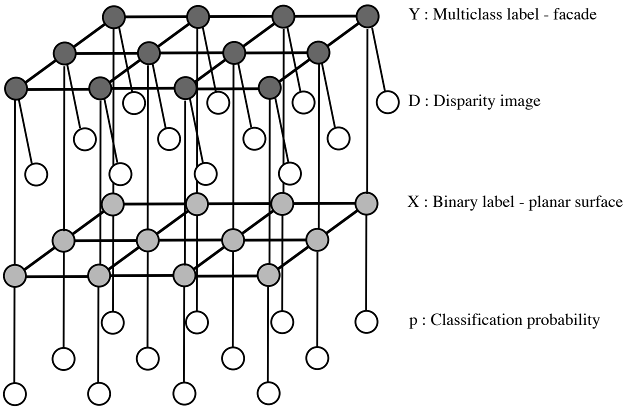

Y : Multiclass label - facade
D : Disparity image
X : Binary label - planar surface
p : Classification probability

Figure 5: Our two-layer MRF model.

4.3.1. Mid-level Representation

We want our energy function for the mid-level model to capture the confidence (probability) of our discriminative classification, and we want there to be a penalty whenever a pixel with a high confidence is mislabeled, but a smaller penalty for pixels with lower confidence in their a priori classification. We use an Ising model to represent our mid-level MRF, where our labels x_s, for $s \in \lambda$ our image lattice, come from the set $\{-1, 1\}$. We define a new variable b_s to represent a mapping of the $X_s \in \{-1, 1\}$ label to the set $\{0, 1\}$ by the transformation $b_s = \frac{X_s+1}{2}$. For a particular configuration of labels l, we define our mid-level energy function as:

$$E(l) = \sum_{s \in \lambda} [(1 - b_s)p(s) + b_s(1 - p(s))] - \gamma_m \sum_{s \sim t} x_s x_t \qquad (18)$$

where $p(s)$ is the discriminative classification probability at s and γ_m is a constant weighting the unary and binary terms. The b_s quantity in the unary term essentially switches between a penalty of $p(s)$ if the label at s is set to -1, and a penalty of $1 - p(s)$ if the label at s is set to 1. Thus for $p(s) = 1$, labeling $x_s = -1$ will incur an energy penalty of 1, but labeling $x_s = 1$ will incur no penalty. Similarly for $p(s) = 0$, labeling $x_s = -1$ will incur no penalty, but labeling it 1 will incur a penalty of 1. A probability of 0.5 will incur an equal penalty with either labeling. Our smoothness term is from the standard Ising model. In our experiments, we used a γ_m value of 10.

4.3.2. *High-level Representation*

In designing our energy function for the high-level MRF, we want to penalize points which are labeled as being on a plane, but which do not fit the corresponding plane equation well. Our set of facade labels y_s, for $s \in \lambda$, is $\{0, \ldots, m\}$, with m equal to the number of candidate planes identified in the plane detection step. It corresponds to the set of candidate planes indexed from 1 to m, as well as the label 0, which corresponds to "not on a plane". We define a set of equations $E_p(s)$ for $p \in \{0, \ldots, m\}$ such that

$$E_p(s) = \mid a'_p u + b'_p v + c'_p - D(s) \mid \tag{19}$$

where the surface normal $\mathbf{n'_p} = (a'_p, b'_p, c'_p)$ corresponds to the plane with label p, and $D(s)$ is the disparity value at s. We normalize this energy function by dividing by the maximum disparity value, in order to scale the maximum energy penalty down to be on the order of 1. For consistency in our notation, we define $E_0(s)$ to be the energy penalty for a label of 0 at s, corresponding to the "not on a plane" classification. We set $E_0(s) = b_s$, such that a labeling of -1 in the mid-level representation results in $b_s = 0$, so there is no penalty for labeling s as "not on a plane". Similarly, when $x_s = 1$, $b_s = 1$, so there is a penalty of 1 to label any of the non-planar pixels as a plane.

To construct our overall energy function for the high-level MRF, we incorporate the exponential of the set of planar energy functions E_p with a delta function, so the energy cost is only for the plane corresponding to the label y_s. Since we cannot compute E_p without a valid disparity value, we use an indicator variable $\chi_D \in \{0, 1\}$ to switch to a constant energy penalty for all planes and the no-plane option, in order to rely strictly on the smoothness term for that pixel's label. For the smoothness term, we use a Potts model, weighted like the mid-level representation with a constant γ_h. In our experiments, though, this value of γ_h was 1. Thus the high-level energy function we are seeking to minimize is:

$$E(l) = \sum_{s\in\lambda} \sum_{p=0}^{m} \delta_{y_s=p} \cdot \exp\left(\chi_D E_p(s)\right) + \gamma_h \sum_{s\sim t} \delta_{y_s=y_t} \tag{20}$$

4.4. *Energy Minimization*

To perform the energy minimization, we use the graph cuts expansion algorithm, specifically the implementation presented in [26]. We perform the minimization in two stages. We first minimize the energy of the mid-level MRF to obtain an approximation to the optimal labeling of planar surface pixels. This step uses prior knowledge from the discriminative classification. Next, we use the mid-level labeling as well as the detected candidate planes as a prior for the high-level MRF, and we use graph cuts again to compute an approximation to that optimal labeling.

5. Experimental Results

We have performed quantitative experiments using our method on a new dataset that consists of 141 grayscale images from the left camera of a stereo imager[2] each with a corresponding 16-bit disparity map that was computed onboard the camera in real time. All images have 500 × 312 resolution and human-annotated ground truth for both binary classification and facade segmentation. The data was collected on a university campus with range of architectural styles, as well as a business district, and is intended to capture a broad range of common urban settings. There are a total of 251 facades represented in the dataset, and for each one, we have computed a gold-standard plane model from its ground truth facade segmentation. There are six images that do not contain any facades, and among the remaining images of the dataset, many feature occlusions and other objects (cars, trees, people, etc.) common to urban settings, so there is an adequate representation of negative samples.

Existing datasets that contained facade images were not adequate for validating our approach, primarily because they contain only optical images and not disparity maps. Even the datasets that are intended for facade segmentation (for example eTRIMS [34]) do not contain individually segmented facades. We are not aware of another publicly available, human-annotated, quantitative stereo building facade dataset, and we believe this new set, which is the first of its kind, can become a benchmark for the community[3].

In all experiments, any parameters of our method's component algorithms were set consistent with the values previously mentioned in the text.

5.1. *Discriminative Modeling*

We performed 6-fold cross-validation with our method (BMA+D), appearance-only BMA, and standard AdaBoost with pixel features (x & y location) and patch-based Haar features. See Table 1 for a pixel-wise quantitative comparison of these models. With the BMA+D classifier, we obtain a 2% increase in accuracy over appearance-only BMA model, and a 6% increase over the standard AdaBoost classifier. We computed the d' statistic for the image-wise performance of all three classifiers and performed a one-tailed student's t-test on this statistic for all pairs of classifiers. Both BMA and BMA+D exhibited statistically significant performance with p-values below 0.1% when compared to AdaBoost. The comparison of BMA+D to appearance-only BMA resulted in a p-value of 8.5%, which, when coupled with the summary statistics in Table 1, indicates at least modest statistical significance to the improvement in classification accuracy. Taken over the entire dataset, these results imply that in this problem domain, disparity features are a beneficial addition to an appearance-only model.

Figure 6 shows ROC curves for these classifiers. Additionally, one image from each validation set was randomly selected for visual comparison of the three methods. Figure 7 shows the probability map of the classifier's output for each of the methods, along with the two-class labeling with a threshold of

[2] Tyzx DeepSea V2 camera with 14 cm baseline and 62° horizontal field of view.

[3] Our dataset is publicly available at: `http://www.cse.buffalo.edu/~jcorso/r/gbs`

Table 1: Quantitative scores for the AdaBoost, BMA, and BMA+D classifiers on the building and background (BG) classes. Recognition rates are computed pixel-wise over the entire dataset.

	True\Pred	BG	Building
ADB	BG	63.58	36.42
BMA		72.73	27.27
BMA+D		**75.33**	**24.67**
ADB	Building	24.67	75.33
BMA		23.97	76.03
BMA+D		**23.51**	**76.49**
ADB	F-scores	0.6876	
BMA		0.7282	
BMA+D		**0.7421**	

Table 2: Recognition rate for the building class on the eTrims 8-class dataset [34]. Note: BMA performs 2-class labeling, all other methods perform 8-class segmentation.

Method	**Rec. Rate (%)**
ICFHGS- [35]	71.9
BMA	70.3
ICF [35]	62.0
RDF-meanshift [36]	60
RDF-watershed [36]	59
ICFwoC [35]	41.1

0.5. Of these six examples, the appearance-only BMA model achieved the best accuracy (2% more than BMA+D) for one image, and the AdaBoost classifier achieved the best accuracy (4% more than BMA+D) for another. However, for the other examples, the BMA+D model outperforms the other classifiers by as much as 8%, and the confidence shown in the probability map is often higher for both classes. Since the probability map acts as a prior for the mid-level MRF labeling, higher confidence from discriminative modeling can translate to higher accuracy in the MRF binary classification.

Although the state-of-the-art in facade segmentation comes as part of multi-class approaches, we compare the two-class BMA approach to the methods in [36, 35] in Table 2 in order to place our results in the context of the existing literature. Since our BMA+D and MRF methods require disparity maps in addition to camera imagery, we are limited to comparison with the appearance-based BMA version. These semantic segmentation methods use the eTrims dataset [34] and label buildings as well as 7 other classes. We performed two-class labeling, an admittedly easier task, on the same dataset, using 40 images for training and 20 for testing as in [35]. But since our goal of facade modeling does not require full semantic segmentation of the scene, we do not extend our approach to the multi-class case. The performance without the inclusion of disparity features or subsequent MRF segmentation is consistent with the labeling accuracy of the building class from the state-of-the-art multi-class labeling approaches.

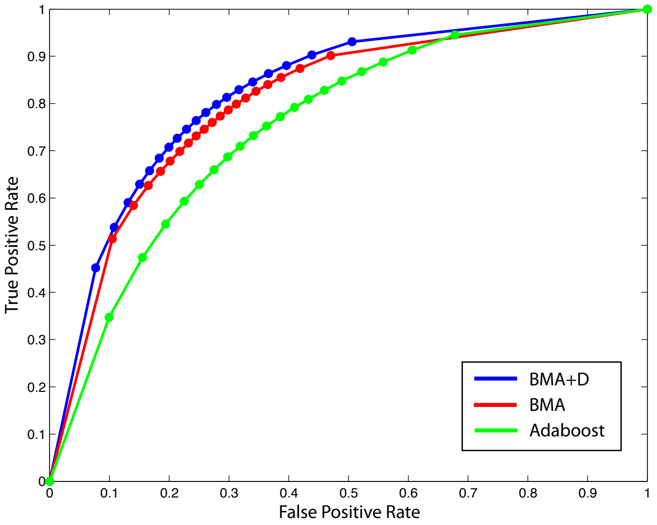


Figure 6: ROC curves for our BMA+D method (blue), appearance-only BMA (red), and patch-based AdaBoost (green).

Table 3: Quantitative scores for the mid-level MRF labeling and the BMA+D classifier on the building and background (BG) classes.

	True \Pred	BG	Building
BMA+D	BG	75.33	24.67
MRF		**79.98**	**20.01**
BMA+D	Building	23.51	76.49
MRF		**21.15**	**78.85**
BMA+D	F-scores	0.7421	
MRF		**0.7773**	

5.2. *Facade Detection*

The mid-level MRF results exhibit further improvement in accuracy over BMA+D alone. Table 3 shows a pixel-wise quantitative comparison of these two methods. With the Bayesian inference of the MRF, we achieve a classification accuracy of almost 80% for each class, and an improvement in overall accuracy of 9% over AdaBoost, 5% over BMA, and 3% over BMA+D.

5.3. *Facade Segmentation and Parameter Estimation*

We computed the facade segmentations and the plane parameters for each of the labeled planes in all of the images from the dataset; some examples are shown in Figure 9. For each of the manually labeled planes in the dataset, we computed ground truth parameters by sampling the labeled region and using RANSAC to determine the plane parameters. Out of 251 total facades in the set, 40 of them were misclassified as background by the mid-level labeling. The other 211 facades were

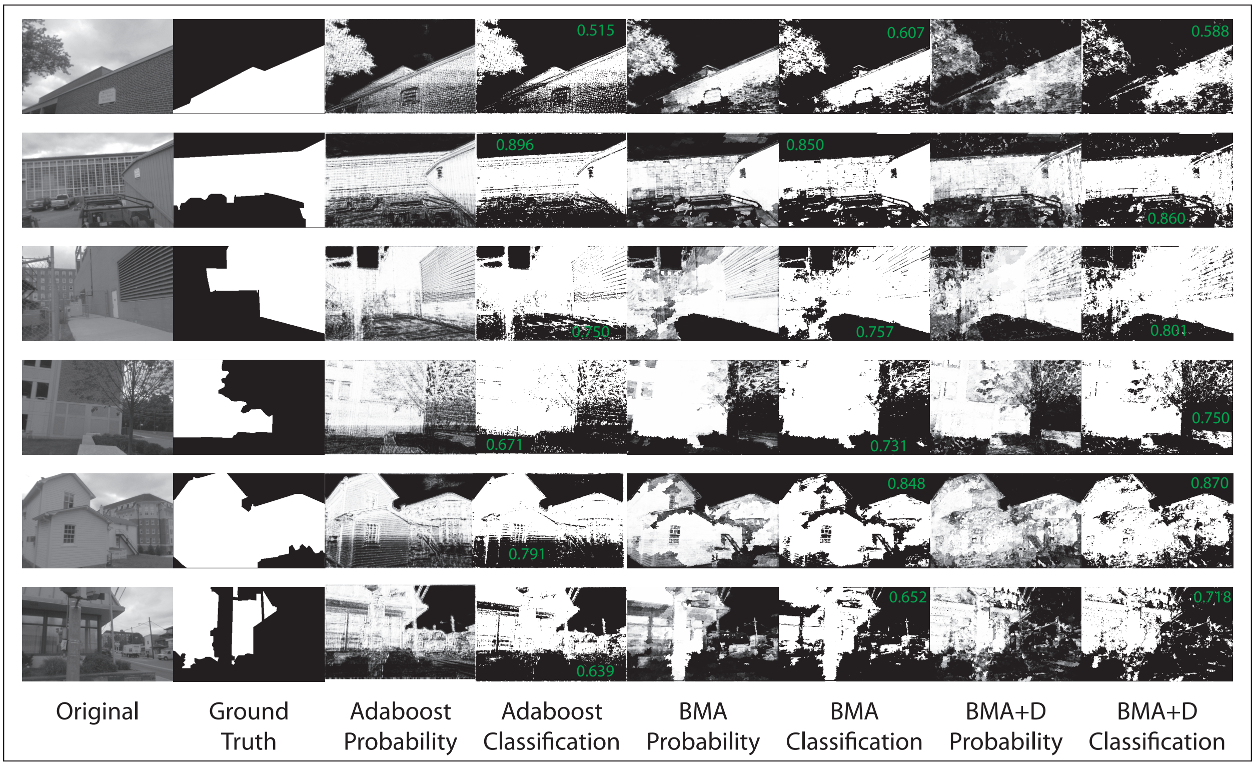


Figure 7: Some examples of discriminative model output. One image was selected at random from each of the 6 validation sets. F-scores are annotated on each classified image.

labeled with at least one candidate plane in the high-level labeling for a detection rate of 84%.

As noted above, some of the ground truth facades are not detected by the mid-level MRF, but multiple segmented planes per ground truth facade are also common. In order to assess the accuracy of our plane parameter estimation, we compute a weighted error measure as the mean pixel-wise angular error between the labeled plane and the ground truth facade, averaged over all pixel in the dataset where the ground truth and high-level labeling are both non-null. Our angular error metric is the dihedral angle between the estimated plane and the ground truth plane (with normal vectors $\mathbf{n}_e$ and $\mathbf{n}_g$, respectively):

$$\phi = \arccos(\mathbf{n}_e \cdot \mathbf{n}_g)$$

The average angular error for any such pixel over the entire dataset is 24.07°. A histogram showing the relative number of pixels labeled with a plane model having angular error in each bin (see Fig. 8) indicates that the peak of the distribution of errors is the range of 0 – 10°. Similarly, the examples shown in Figure 9 indicate that some facades are modeled very accurately, while others have high angular error. This discrepancy motivates our further analysis, which we discuss in the next section.

5.4. Analysis

Our method often segments a detected facade into multiple plane labels, which makes 1-to-1 comparison difficult. In order to overcome this challenge, and to examine the error distribution of Fig. 8 further, we consider two methods for comparing

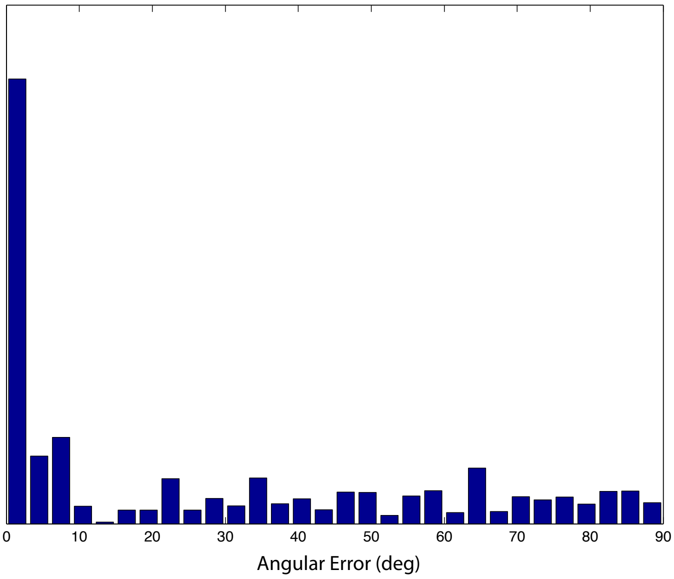


Figure 8: Pixel-wise angular error histogram representing the relative number of pixels that are labeled with a plane model having corresponding angular error across the full dataset .

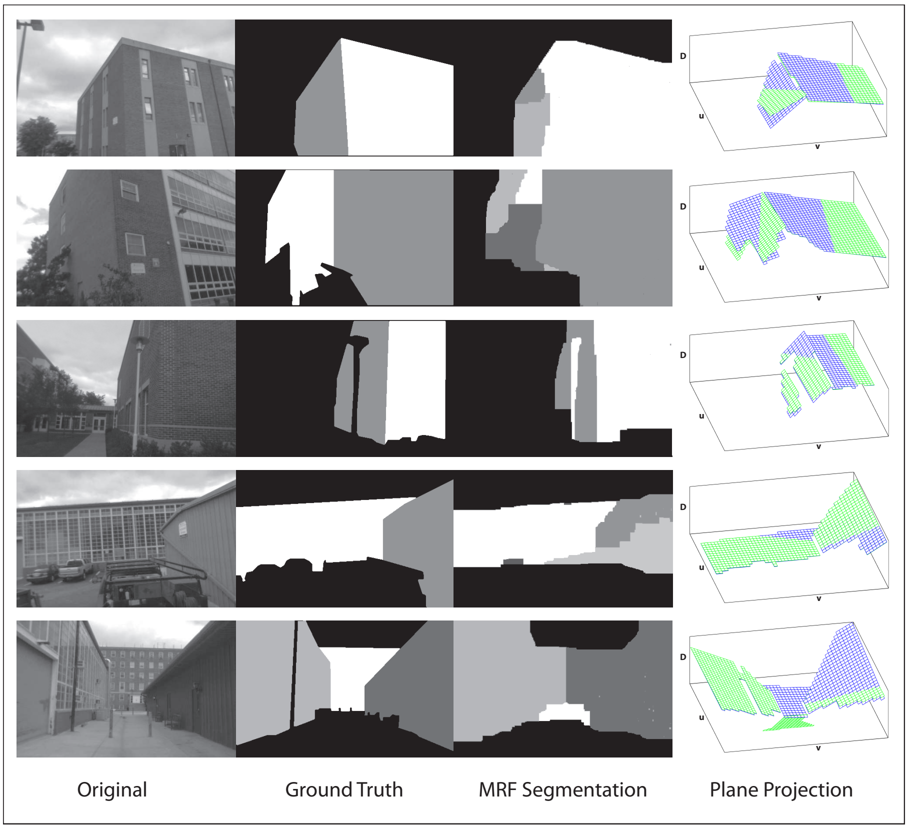


Figure 9: Some examples of MRF labeling output. For each ground truth facade (blue), the closest-fitting plane from the MRF (green) is projected along with it to illustrate the accuracy of the estimation in three dimensions.

the segmentations to the ground truth. First, for each ground truth facade, we compare to it the plane whose label occupies the largest portion of that facade's area in our segmentation. We have noticed that there is often one (or more) accurate plane estimate on each ground truth facade, but it may only cover a minority of the ground truth facade. For example, in the second example of Figure 9, the facade on the left in the ground truth is best modeled by the plane corresponding to the white label in the estimate, but the majority of that facade is labeled with less accurate planes. In order to measure the accuracy of our method in estimating at least some portion of each ground truth facade, our second method of comparison chooses the most accurate plane estimate out of the set of labels that cover each facade's region. In both cases, we compute the average angular error between the chosen segmented plane (largest or best) and the ground truth facade, weighted by the size of the segment, as well as the average percentage of the ground truth facade covered by the chosen label. These results are collected in Table 4. Additionally, for both methods a histogram showing the distribution of chosen labels binned by both angular error and size as a percentage of the frame area can be seen in Fig. 10.

These histograms indicate that most of the high-error segmentations occur with small areas: for both of the methods, the vast majority of facades larger than 10 % of the frame have less than 10 degree error. This implies that the errors are generally small (< 10 %) for the major facades in the image, and it may be possible to restrict or post-process the labeling to eliminate the minor and erroneous plane labels, although that is beyond the scope of this paper.

Table 4: Accuracy for our two methods of comparison to ground truth: largest segment and most accurate segment

Method	Avg. Err.	Avg. Size (% of GT area)
Largest	21.973	66.57
Best	13.765	53.00

The quality of the disparity map is likely to be at least somewhat responsible for this phenomenon, as the usable range of most stereo cameras is limited. For example, the camera used to capture our dataset can only resolve features up to 45 *cm* at a distance of 15 *m*. Thus, even moderately distant facades are likely to be significantly more prone to large errors in their estimates; they will be both small in the frame and less likely to find an accurate consensus set in RANSAC due to the uncertainty in their disparity values. Similarly, for a facade with many invalid disparity values, it may not be sampled adequately, and the points it does have may erroneously be included as part of an inlier set that does not actually lie on the facade. Perhaps on account of this phenomenon, we have observed that many of the high-error segmentations are rotated primarily about a horizontal axis, but are much more accurate in their rotation about a vertical axis. Under the assumption that facades tend to be

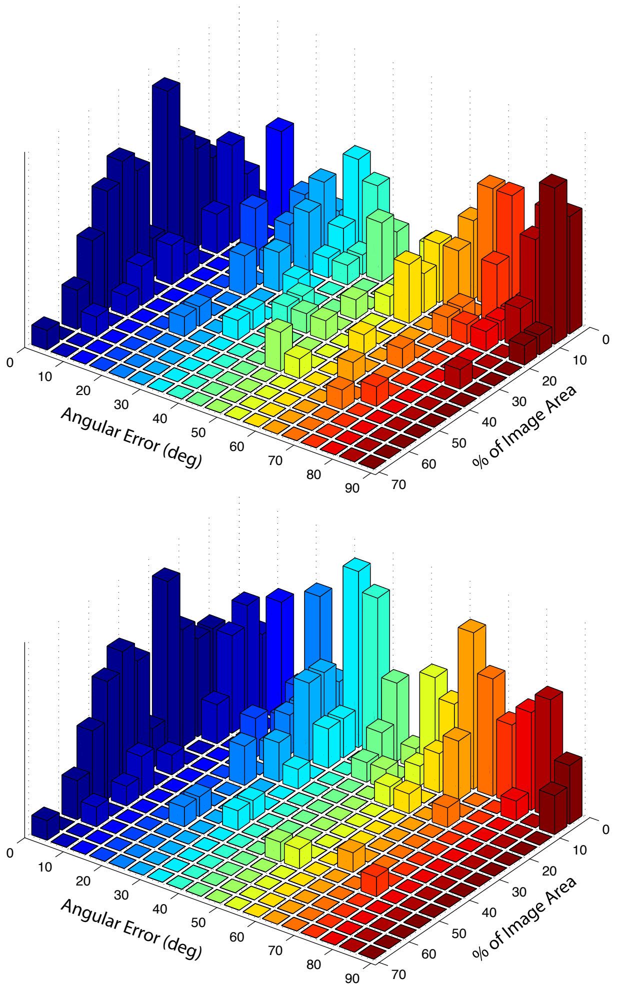


Figure 10: Histogram of angular error per segment, with associated segment size (as a % of the image) for the largest segment (top) and the most accurate segment (bottom). Blue represents smaller error and red represents larger error.

vertical planes, it would be possible to impose a verticality constraint into the RANSAC plane model to restrict the candidate plane set to only vertical plane models.

Without the context of the ground truth facade segmentation, it would not be possible to choose the largest or best label as we do in this analysis, but it is encouraging that on average we are able to achieve < 15% error over a majority of each facade. This result will motivate some of our future work in developing ways to better disambiguate the labels in order to decrease those average errors and increase the area of the most accurate labels.

6. Conclusions

We have presented a system for automatic facade detection, segmentation, and parameter estimation in the domain of stereo-equipped mobile platforms. We have introduced a discriminative model that leverages both appearance and disparity features for improved classification accuracy. From the disparity map, we generate a set of candidate planes using RANSAC with a planar model that also incorporates local PCA estimates of plane normals. We combine these in a two-layer Markov Random Field model that allows for inference on the binary (building/background) labeling at the mid-level, and for segmentation of the identified building pixels into individual planar surfaces corresponding to the candidate plane models determined by RANSAC.

Our BMA+D discriminative model provides superior performance to other classifiers using only appearance features, and our mid-level MRF labeling has proven to further improve the accuracy of the classification to approximately 80%. We were able to identify 84% of the building facades in our dataset, with an average angular error of 24° from the ground truth. However, the distribution of errors peaks in frequency below 10°, indicating that a large percentage of the labels provide very accurate estimates for the ground truth, although some of the labels produced by our method have very high error. Further analysis shows that these high-error labelings most often occur on small segmented regions. Thus our method produces accurate plane estimates for at least the major facades in the image.

A further approach that may enhance these results is strict enforcement of a verticality constraint on the candidate plane models. Extraction of the ground plane would enable us to leverage the assumption that building facades, in general, are perpendicular to the ground plane. Using only locally vertical candidate plane models is an avenue of future work in this area. Another avenue for future investigation is the integration of the distance-based uncertainty of each point in disparity space into the RANSAC models in order to encourage plane fitting to the more accurate points close to the camera. We also intend to pursue other methods for either improving the quality of the input data (e.g. multiview stereo) or improving the methods of compensating for difficult disparity maps.

Acknowledgments

The authors are grateful for the financial support provided in part by NSF CAREER IIS-0845282, DARPA W911NF-10-2-0062, and ARO Young Investigator W911NF-11-1-0090. We are also thankful the support of the United States Army Research Laboratory.

[1] A. Georgiev, P. Allen, Localization methods for a mobile robot in urban environments, IEEE Transactions on Robotics 20 (5) (2004) 851–864.

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

[3] J. Delmerico, J. Corso, P. David, Boosting with Stereo Features for Building Facade Detection on Mobile Platforms, in: e-Proceedings of Western New York Image Processing Workshop, 2010.

[4] Y. Freund, R. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.

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

[6] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[7] J. Delmerico, P. David, J. Corso, Building facade detection, segmentation, and parameter estimation for mobile robot localization and guidance, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2011, pp. 1632–1639.

[8] K. Konolige, M. Agrawal, R. Bolles, C. Cowan, M. Fischler, B. Gerkey, Outdoor mapping and navigation using stereo vision, in: Proceedings of the International Symposium on Experimental Robotics, 2008.

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

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

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

[12] T. Korah, C. Rasmussen, Analysis of building textures for reconstructing partially occluded facades, in: Proceedings of the 10th European Conference on Computer Vision, Springer-Verlag, 2008, pp. 359–372.

[13] B. Frohlich, E. Rodner, J. Denzler, A fast approach for pixelwise labeling of facade images, in: Proceedings of the 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3029–3032.

[14] A. Wendel, M. Donoser, H. Bischof, Unsupervised Facade Segmentation using Repetitive Patterns, Pattern Recognition (2010) 51–60.

[15] D. Hoiem, A. Efros, M. Hebert, Geometric context from a single image, in: Proceedings of the 10th IEEE International Conference on Computer Vision, Vol. 1, IEEE, 2005, pp. 654–661.

[16] M. Recky, A. Wendel, F. Leberl, Façade segmentation in a multi-view scenario, in: Proceedings of International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, IEEE, 2011, pp. 358–365.

[17] J. Xiao, T. Fang, P. Zhao, M. Lhuillier, L. Quan, Image-based street-side city modeling, in: ACM Transactions on Graphics (TOG), Vol. 28, ACM, 2009, p. 114.

[18] A. Yang, S. Rao, A. Wagner, Y. Ma, Segmentation of a piece-wise planar scene from perspective images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2005, pp. 154–161.

[19] A. Flint, D. Murray, I. Reid, Manhattan scene understanding using monocular, stereo, and 3d features, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2228–2235.

[20] I. Posner, D. Schroeter, P. Newman, Describing composite urban workspaces, in: IEEE International Conference on Robotics and Automation, 2007, pp. 4962–4968.

[21] P. David, Detecting Planar Surfaces in Outdoor Urban Environments, Tech. rep., ARMY Research Lab, Adelphi, MD. Computational and Information Sciences Directorate (2008).

[22] J. Bauer, K. Karner, K. Schindler, A. Klaus, C. Zach, Segmentation of building models from dense 3D point-clouds, in: Proceedings of the 27th Workshop of the Austrian Association for Pattern Recognition, Citeseer, 2003, pp. 253–258.

[23] S. Lee, S. Jung, R. Nevatia, Automatic integration of facade textures into 3D building models with a projective geometry based line clustering, Computer Graphics Forum 21 (3) (2002) 511–519.

[24] R. Wang, J. Bach, F. Ferrie, Window detection from mobile LiDAR data, in: Proceedings of the IEEE Workshop on Applications of Computer Vision, IEEE, 2011, pp. 58–65.

[25] D. Gallup, J. Frahm, M. Pollefeys, Piecewise planar and non-planar stereo for urban scene reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1418–1425.

[26] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2002) 1222–1239.

[27] J. Kosecka, W. Zhang, Extraction, matching, and pose recovery based on dominant rectangular structures, Computer Vision and Image Understanding 100 (3) (2005) 274–293.

[28] W. Zhang, J. Kosecka, Image Based Localization in Urban Environments, in: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission, IEEE Computer Society, 2006, pp. 33–40.

[29] D. Robertson, R. Cipolla, An image-based system for urban navigation, in: Proceedings of the British Machine Vision Conference, Vol. 1, Citeseer, 2004, pp. 260–272.

[30] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (1) (2002) 7–42.

[31] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1297 –1304.

[32] V. Kolmogorov, Convergent tree-reweighted message passing for energy minimization, IEEE Transactions on Pattern Analysis and Machine Intelligence (2006) 1568–1583.

[33] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, W. Stuetzle, Surface reconstruction from unorganized points, Computer Graphics 26 (2) (1992) 71–78.

[34] F. Korč, W. Förstner, eTRIMS Image Database for interpreting images of man-made scenes, Tech. Rep. TR-IGG-P-2009-01, Dept. of Photogrammetry, University of Bonn (April 2009).
URL `http://www.ipb.uni-bonn.de/projects/etrims_db/`

[35] B. Fröhlich, E. Rodner, J. Denzler, Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach, in: Asian Conference on Computer Vision (ACCV), Springer, 2012, pp. 218–231.

[36] M. Y. Yang, W. Förstner, D. Chai, Feature evaluation for building facade images an empirical study, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B3.