Using Stereo for Object Recognition

Scott Helmer and David Lowe

***Abstract*— There has been significant progress recently in object recognition research, but many of the current approaches still fail for object classes with few distinctive features, and in settings with significant clutter and viewpoint variance. One such setting is visual search in mobile robotics, where tasks such as finding a mug or stapler require robust recognition. The focus of this paper is on integrating stereo vision with appearance based recognition to increase accuracy and efficiency. We propose a model that utilizes a chamfer-type silhouette classifier which is weighted by a prior on scale, which is robust to missing stereo depth information. Our approach is validated on a set of challenging indoor scenes containing mugs and shoes, where we find that priors remove a significant number of false positives, improving the average precision by 0.2 on each dataset. We additionally experiment with an additional classifer by Felzenszwalb *et al.*[1] to demonstrate the approach's robustness.**

I. Introduction

Object classification and recognition has progressed rapidly in recent years due to advances in machine learning, more sophisticated feature extraction techniques, and the ever greater availability of image datasets. Despite the recent success, there is still significant progress required before we have robots assisting the elderly, cleaning our homes, or fetching household items. A particular challenge for mobile robots in an indoor environment is that most of the objects to be manipulated occupy small portions of cluttered scenes. However, much of the success thus far in object recognition/localization has been achieved with large objects that are often assumed to occupy a significant portion of the image, such as pedestrians, vehicles, and animals [2]. Many smaller objects found within cluttered indoor scenes are left relatively unaccounted for, such as mugs, staplers, shoes, etc. Due to large variations in appearance, these types of object categories are difficult to recognize with patch based methods, which generally require significant resolution and distinctive features that are internal to the object and therefore not disturbed by background clutter.

There is an increasing body of work suggesting that using scene context can improve the efficiency and accuracy of localization. Scene cues such as the gist of a scene were successfully used by Torralba [3] to predict the likely location and scale of objects. It has also been shown that local context, such as surrounding image texture can improve object classification and segmentation [4]. Object co-occurrence and co-location have also been shown by many researchers to be valuable in object detection [5], [6], [7]. More recently, many research results have demonstrated that using 3D scene information such as surface orientation and scale [8], [9], [10], [11], [12], [13] can be useful in object recognition. In the case where the objects in a scene lack sufficient resolution, or are difficult to detect with current recognition methods, scene context can play a large role.

The focus in this paper is on fusing 2D image information and depth information from stereo images into one model for localization, particularly in the case of contour-based objects. A primary motivation for our work stems from our recent experiences in designing a vision system for our robot Curious George [14], an entrant in the Semantic Robot Vision Challenge (SRVC). This contest is a visual scavenger hunt, where a robot explores a room and returns a set of images corresponding to a list of objects it was tasked to find. Despite winning the competition in 2007 and 2008, it was apparent that recognition of generic objects from arbitrary viewpoints is still very much an open challenge. With stereo vision, we can use a prior on object size, which can be as simple as a mean and variance of an object's real world size. We show that this reduces false positives, and can increase computational efficiency, which is particularly important with a mobile robot. In conjunction with the prior, we experiment with methods to utilize surface variation with a contour-based classifier. The primary contribution of this paper is a model formulation that is robust to missing information from stereo. We validate our approach on a challenging set of scenes containing shoes and mugs.

A. Related Work

There are a variety of recent approaches that make use of an object's real world scale as a prior for the scale in the image. One of the pioneering works in this regard, and the most similar to our formulation, is that of Hoiem *et al.*[8]. Using only an image, the approach jointly infers 3D object locations and scene information, such as 3D surface orientation, ground plane, and horizon. Assumptions using the estimated horizon and a prior on an object's real scale are utilized to provide a prior on the expected scale in the image. This prior modulates the response from an appearance-based object detector, showing marked improvement in pedestrian and vehicle localization. A primary distinction between our work and theirs is that our scale prior uses stereo rather than their horizon assumptions, which are not applicable indoors.

Another approach that is similar in spirit to our own is the pedestrian detection work of Gavrila and Munder [15]. Here, they utilize sparse stereo and ground plane constraints to determine regions of interest, where they will then run a detector for a pedestrian at the appropriate scale. Similar

Scott Helmer and David Lowe are with the University of British Columbia, Department of Computer Science, Vancouver, British Columbia, Canada, shelmer@cs.ubc.ca

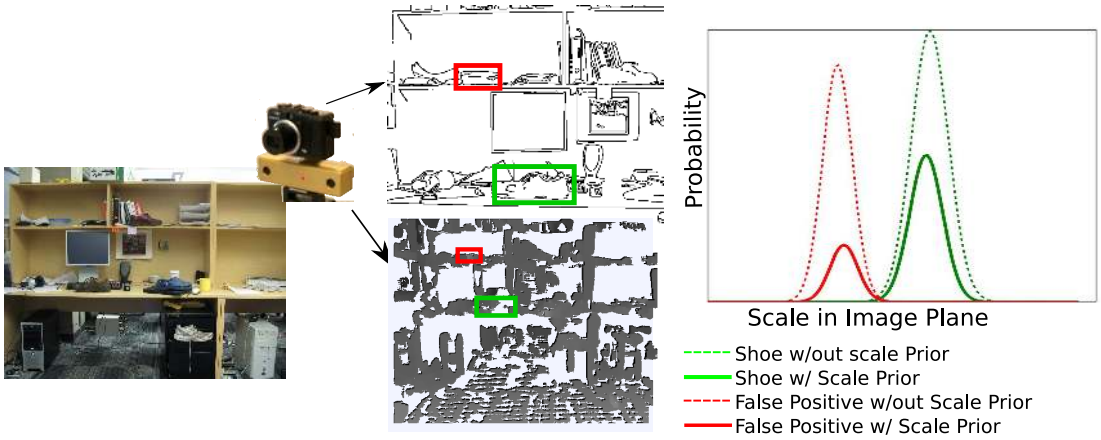


Fig. 1. A stereo camera and monocular camera produce an edge image and a depth map. Missing depth information is shown in white. A contour-based classifier evaluates bounding boxes in the edge image. The scale prior weights these scores depending on whether the depth and scale of the bounding box agree, thereby reducing the score of false positives to a much greater degree than true positives.

to our own work, they also make use of a chamfer distance metric. One of the primary distinctions between our work and their's is that our approach does not require ground plane constraints, and is possibly more robust to missing stereo information. Moreover, our approach follows a clean probablistic framework, where as their approach involves numerous parameters and thresholds that requires extensive training to set. Other recent approaches using object scale in recognition are those of Gould *et al.*[10], and Quigley *et al.*[13]. They utilize a mobile robot to acquire high-accuracy depth maps using a laser scanner, which requires 2-3 seconds per scan. From this data, their object detector utilizes both surface variation, 3D shape, and appearance to find objects such as mugs, cups, staplers, etc. Their system achieves impressive results, but their use of a laser scanner to acquire data is unrealistic for many applications. Our approach utilizes stereo which is faster, cheaper, and less invasive but also requires additional robustness to uncertainty.

In regards to object detection, most successful approaches have utilized patch based techniques that decompose recognition into recognizing parts of the object. These range from bag-of-word approaches that discard 2D spatial relationships entirely, to approaches that model the spatial relation between features. However, for classes that are visually defined by their 3D shape, the signal from the identifying contour on a patch is sometimes overwhelmed by the noise of foreground texture and background clutter. There are numerous recent localization approaches, however, that make use of contours and chamfer matching, including [16], [17], [18], [19], [20]. We also experiment with extensions to chamfer matching that utilize depth information.

II. Method

The task we are concerned with is localizing an object, obj, in an intensity image I_m, while also leveraging a noisy depth image I_z. To achieve this we adopt and adapt a multi-scale sliding window classifier, a widely used approach to localization. Here, a subset of N windows, $\{\theta_i\}_{i=1..N}$, are evaluated to determine a score as to whether they contain the object of interest. We achieve this by using a probability function $p(\mathbf{o}, \theta | I_z, I_m)$, where $\mathbf{o} = \{obj, background\}$. Finally, the scores from the sub windows are combined using non-maximum suppression to determine likely detections.

There are two directions from which we can improve object localization: reducing false positives, and increasing the scores for true positives. Given that depth images derived from stereo data are noisy and that the objects are small relative to the depths involved, we cannot place much hope in classification based solely on the depth data. Instead, we focus our efforts on utilizing the depth data to reduce false positives, and to help make the most use of the appearance information.

To achieve this we separate information about the appearance, the scale of a scene element (obj or $background$), and the surface variation. We first assume that the appearance, I_m, and depth information, I_z, are conditionally independent given $\mathbf{o}$. This is not technically true since depth using stereo is derived from appearance information, the effect of this dependence is minimal for recognition. Next, the bounding box θ implies a centre, scale s_θ, and aspect ratio (which we assume is fixed), and we denote $I_z(\theta)$ as the depth values within θ. If we assume that a scene element's presence and appearance are independent of where it is in the scene, then the nature of the element's surface is independent of where it is in the scene. So, if we have a scalar function $f_\theta(I_z)$ that measures where the surface of the element is in the scene, then we can move that surface to the model coordinate system, $I_v(\theta) = I_z(\theta) - f_\theta(I_z)$. We discuss $f_\theta(I_z)$ further in Section II-A. As stated, now $\{f_\theta(I_z), I_v(\theta), I_m\}$ are statistically independent. So, at a particular bounding box, $\{f_\theta(I_z), I_v(\theta)\}$ become our features that describe the depths in that bounding box. We define our score as the probability,

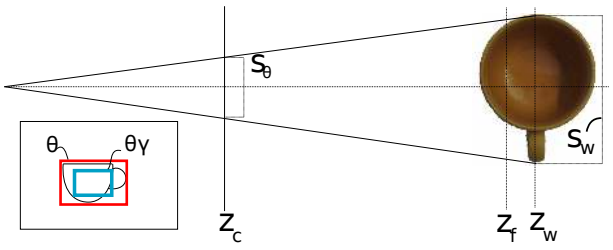


Fig. 2. Scene Geometry

$$p(\mathbf{o},\theta|I_z,I_m) = \frac{p(f_\theta(I_z),I_v(\theta),I_m|\mathbf{o},\theta)p(\mathbf{o},\theta)}{p(I_z,I_m)} \quad (1)$$

$$= \frac{p(\mathbf{o}|I_m,I_v(\theta),\theta)p(f_\theta(I_z)|\mathbf{o},\theta)p(\theta,I_v(\theta),I_m)}{p(I_z,I_m)} \quad (2)$$

$$\propto p(\mathbf{o}|I_m,I_v(\theta),\theta)p(f_\theta(I_z)|\mathbf{o},\theta) \quad (3)$$

where Equation 2 follows 1 by independence and an application of Bayes rule. The final equation follows because we assume uniformity in terms not involving $\mathbf{o}$.

This formulation is similar in spirit to that of Hoiem *et al.*[8], and likewise is general and not dependent upon our choice of classifier and priors. The first term is the object classifier which we describe in Section II-B. However, note that the classifier can depend on both surface data in $I_v(\theta)$ and the intensity image I_m, which allows for a more powerful classifier based upon both shape and texture. The second term is related to our prior on the scale of a scene element With both these terms, we are primarily concerned with detecting the object of interest, obj, so we only search through our $p(\mathbf{o},\theta|I_z,I_m)$, for when $\mathbf{o} = obj$. It's necessary to include $\mathbf{o}$ in the formulation so that we can utilize the object classifier $p(\mathbf{o}|I_m,I_v(\theta),\theta)$.

A. Object Scale from Depth

The scale prior of the object is captured in the distribution $p(f_\theta(I_z)|\mathbf{o},\theta)$, which should also capture the uncertainty in depth measurements. The scale of an object class could be arbitrarily complex, as could the mechanism that is responsible for errors in depth measurements. In this section we first formulate the prior generically, and then describe the approximations made in our approach.

The geometry of a scene is illustrated in Figure 2. Formally, z_w is the distance of the object's centre to the camera's origin, s_θ is the scale (i.e., height or width) in the image plane, and s_w is the scale of the object in the fronto-parallel plane at z_w. For the moment we will restrict the object to being represented as a plane parallel to the camera. We denote the focal length of the camera, z_c, and the baseline distance between stereo cameras as b. The disparity d_x on the image plane for any point $x \in \theta$ is d (since the object is a plane), and the number of points in the bounding box is N. Using perspective projection and epipolar geometry, the following relationships hold

$$z_w = \frac{z_c s_w}{s_\theta} = \frac{z_c b}{d} \Rightarrow d = \frac{b s_\theta}{s_w} \quad (4)$$

Now, in practice we do not know the true values of z_w, s_w, or d, so we use instead the random variables $\mathbf{z}$, $\mathbf{s}$, $\mathbf{d}$. For a particular point x, $\mathbf{d}_x = \mathbf{d} + \boldsymbol{\xi}_x$, where $\boldsymbol{\xi}_x$ is noise due to discretization and the stereo algorithm. We can use the relationships in equation 4 to show,

$$\frac{1}{N}\sum_{x\in\theta}\mathbf{z}_x = \frac{z_c b \mathbf{s}}{s_\theta b + \mathbf{s}\frac{1}{N}\sum_x \boldsymbol{\xi}_x} \quad (5)$$

$$\approx \frac{z_c \mathbf{s}}{s_\theta} \quad (6)$$

We assume the errors, $\{\xi_x\}_{x\in\theta}$, are zero mean and independent, and by the central limit theorem the second term in denominator becomes 0, allowing the approximation in equation 6. Using this, we define $f_\theta(I_z)$ as the average depth in an area within our bounding box, so $f_\theta(I_z) \sim z_c\mathbf{s}/s_\theta$. The average depth is taken over a bounding box that is γ times the size of the original bounding box θ, but centered at the same point in the image, as in Figure 2, a technique also adopted by [13]. The reason for this is that depth values around the edges of θ will be more likely to fall on the background. In our experiments we use $\gamma = 0.8$. Formally, if we denote θ_γ as this inner bounding box of θ, and missing depth values as $\oslash$, then

$$f_\theta(I_z) = mean(\{I_z(x)|x \in \theta_\gamma, x \neq \oslash\}) \quad (7)$$

The prior for the object's scale can take any form. In this paper we assume that the scale of the object class is Gaussian, with parameters $\{\mu_s,\sigma_s\}$. This implies that $f_\theta(I_z)$, ie. $p(f_\theta(I_z)|\mathbf{o},\theta)$, is also a Gaussian with parameters $\{z_c\mu_s/s_\theta, z_c\sigma_s/s_\theta\}$. It should be noted that if we simply tried to use the depth value at the centre of the bounding box, versus an average, the classifier barely outperformed the base classifier, in part because stereo often gives no reliable values at the centre of a textureless object.

Up until this point we have treated the object as planar. This assumption can be relaxed if the object scales isometrically and we know the aspect ratio of the object's dimensions. Here, the distance between the object's centre plane, z_w, and the frontal plane, z_f, will be proportional to the scale for the object, i.e., $z_w - z_f = \beta s_w$. This amounts to little more than adding $\beta\mu_s$ to the mean for $f_\theta(I_z)$. In practice, this can be ignored for smaller objects.

B. Object Classifier

In this section we describe our object classifier, which provides $p(\mathbf{o}|I_m,I_v(\theta),\theta)$ for equation 3. The object classifier is based upon the insight that for many manufactured objects, there are often no distinctive textural features. As a result, we instead base our object classifier upon the silhouette of the object since this captures succinctly the shape of the object. The measure that we utilize to determine how close the image $I_m(\theta)$ is to the object class is based upon an altered

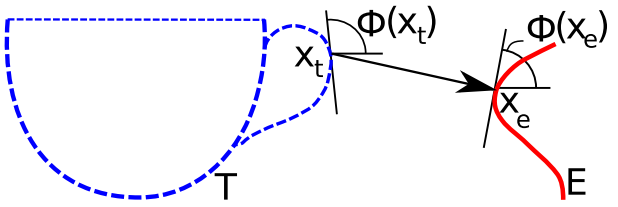


Fig. 3. Similarity between the model silhouette T and the edge image E is based upon the sum of the differences $\|\phi(x_t) - \phi(x_e)\|$, and $\|x_t - x_e\|_2$ for edgels in $x_t \in T$ and their closest matches $x_e \in E$.

version of chamfer distance. Although this is not a state of-the-art classifier, it is surprisingly effective for classifying the canonical viewpoints of some contour-based classes. The classifier we initially describe here does not utilize the shape information contained in $I_v(\theta)$, we discuss an altered version that does near the end of the section.

The chamfer distance was first introduced by [21] as a means of measuring the distance between two curves. In its most basic form, it is the total distance from all the points in a template point set T to the closest points from point set E. A threshold τ is used to account for missing edgels. Relative to some translation $\mathbf{x}$ applied to T, the thresholded chamfer distance is defined as,

$$d^{T,E}_{cham,\tau}(\mathbf{x}) = \frac{1}{|T|} \sum_{\mathbf{x}_t \in T} \min(\tau, \underset{\mathbf{x}_e \in E}{\operatorname{argmin}} \|(\mathbf{x}_t + \mathbf{x}) - \mathbf{x}_e\|_2) \tag{8}$$

This measure however is inherently biased towards cluttered images [18], where a high density of edgels is likely to have a small chamfer distance despite the fact that the pattern of edgels looks nothing like template T. To overcome this, we take an approach similar to Shotton *et al.*[16], who added the difference in orientations between matching edgels to the chamfer distance. Explicitly, if we denote $\mathbf{x}_e^{\mathbf{x_t}+\mathbf{x}}$ to be the $\mathbf{x}_e \in E$ closest to $\mathbf{x}_t + \mathbf{x}$, then we can define two disjoint sets, T' and $\overline{T'}$, where $T' = \{\mathbf{x}_t | \|\mathbf{x}_e^{\mathbf{x_t}+\mathbf{x}} - x_t\|_2 < \tau\}$ and $\overline{T}' = \{\mathbf{x}_t | \|\mathbf{x}_e^{\mathbf{x_t}+\mathbf{x}} - \mathbf{x}_t\|_2 \geq \tau\}$. These sets denote nothing more than splitting T into those edgels with a match less than τ and those that are too far away. If we let to $\phi(\mathbf{x})$ be the orientation of an edgel modulo π, then the orientation penalty is defined as,

$$d^{T,E}_{orient,\tau}(\mathbf{x}) = \frac{2}{\pi|T|}\left(|\overline{T}'| + \sum_{\mathbf{x}_t \in T'} |\phi(\mathbf{x_t}) - \phi(\mathbf{x}_e^{\mathbf{x_t}+\mathbf{x}})|\right) \tag{9}$$

and the total oriented chamfer score is,

$$d^{T,E}_{\tau}(\mathbf{x}) = (1-\lambda) d_{cham,\tau} + \lambda d_{orient,\tau} \tag{10}$$

where λ weights the contribution of the orientation difference to the chamfer score. The Figure 3 sheds some light on the oriented chamfer distance. This description thus far has not touched upon the issue of scale. Again, we follow Shotton *et al.*[16] who scale the template rather than edge image. For a bounding box $\theta = \{s_\theta, \ell_c\}$, we scale the edgels in the template by $s = s_\theta / s_m$, where s_m is the model scale.

Using the chamfer score, $d^{sT,E}_{s\tau}(\mathbf{x})$, in a logistic function, our base classifier becomes

$$p(o|I_m) = [1 + \mathbf{exp}(\alpha_o - \alpha_1 d^{sT,I_e}_{s\tau}(x_\theta))]^{-1} \tag{11}$$

This classifier does not make use of the shape information that is available in the depth variances $I_v(\theta)$. We did implement a variation on the score in equation 10, where we added an additional term that penalized deviations in depths for the matching edgels. The intuition here is that since $I_v(\theta)$ has mean 0, we expect the variation from zero to be small relative to the object's size. We discuss results with this enhanced version of the classifier in the results section as well.

There are a number of relevant parameters of the object classifier that need to be set or learned. λ, which modulates the influence of orientation differences on the distance, was set to 0.25. The parameter τ was set to 0.15. Both of these parameters are similar to values used by Shotton *et al.*[16], and were set by cross-validation on an independent dataset. The object silhouette we utilized was acquired taking an image of a prototypical object on an uncluttered background. From this, we extracted the silhouette by using only the contours on the exterior of the object. The parameters α_o and α_1 of the logistic classifer can be learned using maximum likelihood on training data. For the class of mugs, we used Graz 17 [22], and for the shoes we collected a set of training images from the internet.

C. Sampling for Detection

Our object detector is based upon determining local maxima in the probability function $p(\mathbf{o}, \theta | I_z, I_m)$, which amounts to finding the bounding boxes with high scores. In a multi-scale sliding window setting, this can be computationally expensive depending upon : 1) the computation required to evaluate a single bounding box, and 2) the sensitivity of the classifier to minor changes in scale and location. Computation of a single bounding box is relatively efficient. We use integral images to compute the scale prior and use the distance transform so that chamfer matching is $O(k)$ where k is the number of edgels in our silhouette.

In general, the less sensitive to scale and location a classifier is, the more sparsely we can sample $p(\mathbf{o}, \theta | I_z, I_m)$ and still hope to find all detections. For location, a large portion of the image can be sampled sparsely since the scale prior and chamfer matching both vary somewhat smoothly. Scale sampling is more tricky, since there can be considerable performance degradation if too few scales are evaluated. For chamfer matching, without scale priors, we found that results were stable when sampling every 1/8 octave in scale space, i.e., rescaling by $\{2^{i/8} | i \in \mathbb{I}, -16 < i < 16\}$. This can be reduced to a sampling rate of 1/4 octave and only sampling at a greater frequency in regions where the chamfer distance is small. However, with a full detector that utilizes the scale prior, the number of samples can be greatly reduced. For example, for shoe detection at a particular location, we

could use depth to infer at what scales to sample, allowing a reduction of samples by up to 80 percent.

III. Evaluation

To validate our approach we collected a dataset of stereo and still images for a variety of scenes containing mugs and shoes, examples of which can be found in Figure 5. Using these images, we then compare the performance of our chamfer-distance object detector versus the performance when this base detector is augmented with a prior on scale.

A. Experimental Setup

The intent of our data collection was to produce a set of images that would be challenging for an appearance based classifier due to the significant amount of clutter and textural variation on the object itself. With this in mind, the objects were placed at a variety of depths (1 to 7.5 m), with varying amounts of background clutter. In addition, the shapes and texture of the objects themselves varied. Due to the restricted range of viewpoints covered by our shape model, the objects were placed parallel to the image plane, although the object could be left or right facing. For the mug dataset we collected 20 images from different indoor scenes, with 15 different mugs, with about 3 mugs per scene. For the shoe dataset we also collected 20 images of different scenes, with 8 different shoes, with about 3 shoes per scene.

The camera setup consisted of a Canon G7, using 1216x912 images, and a Bumblebee 2 stereo camera, using 1024x768 images, with the Canon camera on top of the Bumblebee as in Figure 1. The stereo algorithm used was Point Grey's stereo algorithm provided with the camera, which provides fast, accurate depth maps for textured regions, and annotates ambigous regions as missing information. The motivation for the two camera approach is that the quality of the edge images from the Bumblebee camera were poor in comparison to those of the Canon camera. However, this introduces an additional complication since the 3D point cloud derived from the Bumblebee and its software are in the Bumblebee's coordinate system. To overcome this we find a set of point correspondences between one Bumblebee image and the Canon image using SIFT features and geometric constraints [23]. Using the 3D points from the Bumblebee, we fit a projection matrix P, using the Gold standard algorithm [24], that maps all 3D points from the Bumblebee to the Canon. Although this introduces additional errors into the depth image I_z, in practice this produced considerable improvements in both the base classifier and the classifier that utilized scale as a prior. In the case of shoes for example, the average precision for the base detector on Canon images was 0.49, whereas for the Bumblebee images average precision was 0.35.

B. Results and Analysis

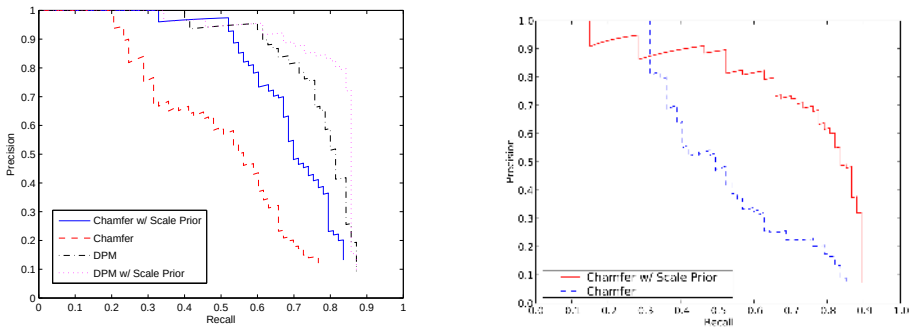
In order to evaluate our approach, we perform detection over a set of scenes, where any bounding box returned by the system is considered a true positive only if it's overlap with the true bounding box is at least 50 percent the area of the union of the two bounding boxes, which is standard for object detection [2]. The primary metric we utilize for comparison is the average precision (AP). For the shoe dataset, the base classifier achieved an AP of 0.51, and an AP of 0.71 with the scale prior. For the mug dataset, the base classifier achieved an AP of 0.48 and an AP of 0.72 with the scale prior. Also, as we can see in the recall precision curves in Figure 4, there is a significant difference in performance in these two classifiers.

We also experimented with an additional classifer, the deformable parts model (DPM) developed by Felzenswalb *et al.*[1], which has source code available on the net. In this case, we only trained the classifier for mugs since it also coincided with our work for SRVC. Using the output from this classifier for $p(\mathbf{o}\|I_m)$, we also found that the results improved with the use of the scale prior on the mugs data set. This can be seen in Figure 4(a), for the DPM models. This demonstrates that the approach is usefull for more than the classifier we outlined earlier. The improvement is not as drastic since the DPM is more sophisticated and trained on a large set of training data.

There are two types of errors made by the object classifier, false negatives and false positives. As can be seen by the shape of the recall-precision curves and from the examples, most of the improvement is a result of fewer false positives. In both object classes, there were a number of instances of false negatives that were due to the failure of the object classifier, irregardless of the scale prior. These failures were due partially to failure in edge detection, but also due to the fact that a simple silhouette does not capture object class variation particularly well.

We also performed a number of experiments to determine the sensitivity of the approach to other parameters. The first set of experiments were conducted in regards to the interplay between the parameter settings for the object classifier, via α_o and α_1, and scale prior σ_s^2. The parameters in the final results were set by optimizing for the α's on a separate dataset, and simply setting the variance, σ_s^2, to an approximation of the real scale variance of the objects. In subsequent experiments we noticed that the results were fairly robust to alterations in these parameters. For example, if we doubled σ_s, the difference in AP were not significant. This suggests that the scale prior is primarily removing egregious false positives, not assisting in modeling fine distinctions between detections that have a scale that agrees roughly with depth.

We also experimented with utilizing the surface variance, $I_v(\theta)$, i.e., the residuals after the mean in the region was subtracted off, to improve detection. In general, for small objects we do not expect a great deal of surface variation, so discontinuities or cases where the template contour is matching a background segment can be detected using $I_v(\theta)$. In one experiment we embedded $I_v(\theta)$ into chamfer matching as mentioned in Section II-B. In another experiment we used a uniform prior on the variation of $I_v(\theta)$, with a range of 0 to object scale. This has the effect of disallowing large discontinuities in depth with the region. In both cases, these approaches did improves results by about 0.025 in AP for



(a) Mugs (b) Shoes

Fig. 4. Recall Precision Curves

both datasets. Again this slight improvement is due to the fact that the scale prior removed a majority of false positives.

C. Conclusion and Future Work

In this paper we have presented an approach that fuses appearance-based recognition using contours with depth information acquired from stereo. Although previous approaches have made use of depth information for recognition, it has yet to be demonstrated that it is feasible with realistic stereo data in an environment not dependent upon ground plane assumptions, in which noise and missing values can be significant. As our results indicate, a prior on scale can be utilized to increase the accuracy, efficiency, and robustness of object localization on the types of objects expected to be manipulated by mobile robots in indoor environments.

In addition, the approach we presented is general in that any object classifier can be used. The limitations of a classifier based upon an entire silhouette are significant, including sensitivity to viewpoint, clutter edgels, and intra-class variation in the shape of the object. Future work will focus on utilizing a more sophisticated object classifier. In the context of a mobile robot, such as a setting like the SRVC, more sophisticated object class models require more computation, making the use of scale as prior that much more important in focusing attention on relevant regions.

Moreover, for shape based objects, the primarily challenges are in disambiguating foreground contours from clutter and modelling variation. Using a scale prior that is independent of the detector can help in reducing false positives, as shown by our results, but provides little help in reducing the noise introduced by clutter edges. Future work will also investigate utilizing depth information to both improve edge detection and assist in reducing the effect of clutter edges on measuring the distance between two curves.

REFERENCES

[1] P. F. Felzenszwalb and D. A. M. andDeva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[3] A. Torralba, "Contextual priming for object detection," *IJCV*, vol. 53(2), pp. 169–191, 2003.

[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition," in *ECCV*, 2006.

[5] A. Torralba, K. Murphy, and W. Freeman., "Contextual models for object detection using boosted random fields," in *NIPS*, 2005.

[6] D. Parikh, L. Zitnick, and T. Chen., "From appearance to context-based recognition: Dense labeling in small images," in *CVPR*, 2008.

[7] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classification," in *ICCV*, 2005.

[8] D. Hoiem, A. Efros, and M. Heber, "Putting objects in perspective," in *CVPR*, 2006.

[9] B. Leibe, N. Cornelis, K. Cornelis, and L. Gool, "Dynamic 3d scene analysis from a moving vehicle," in *CVPR*, 2007.

[10] S. Gould, P. Baumstarck, M. Quigley, , A. Y. Ng, and D. Koller, "Integrating visual and range data for robotic object detection," in *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, 2008.

[11] G. J. Brostow, J. Shotton, J. Fauqueur, , and R. Cipolla, "Segmentation and recognition using structure from motion cues," in *ECCV*, 2008.

[12] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Depth from familiar objects: A hierarchical model for 3d scenes," in *CVPR*, 2006.

[13] M. Quigley, S. Batra, S. Gould, E. Klingbeil, Q. Le, A. Wellman, and A. Y. Ng, "High-accuracy 3d sensing for mobile manipulation: Improving object detection and door opening," in *International Conference on Robotics and Automation*, 2009.

[14] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. A. Baumann, J. J. Little, and D. G. Lowe, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, 2008.

[15] S. M. Dariu M. Gavrila, "Multi-cue pedestrian detection and tracking from a moving vehicle," *IJCV*, vol. 73 (1), pp. 41–59, 2007.

[16] J. Shotton and R. Blake, A.and Cipolla, "Multi-scale categorical object recognition using contour fragments," *PAMI*, 2007.

[17] A. Opelt, A. Pinz, and A. Zisserman, "A boundary fragment model for object detection," in *ECCV*, 2006.

[18] D. M. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," *PAMI*, vol. 29 (8), pp. 1408–1421, 2007.

[19] M. P. Kumar, P. Torr, and A. Zisserman, "Extending pictorial structures for object recognition," in *BMVC*, 2004.

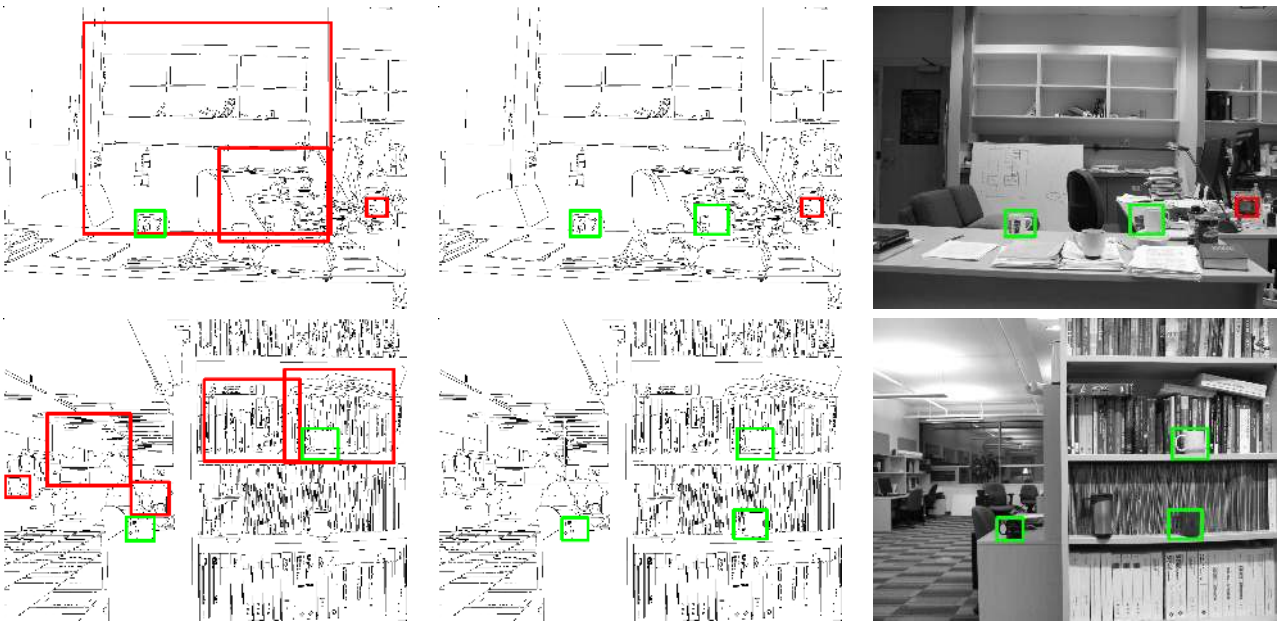
[20] B. Leibe, E. Seeman, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR*, 2005.

[21] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf., "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *IJCAI*, 1977.

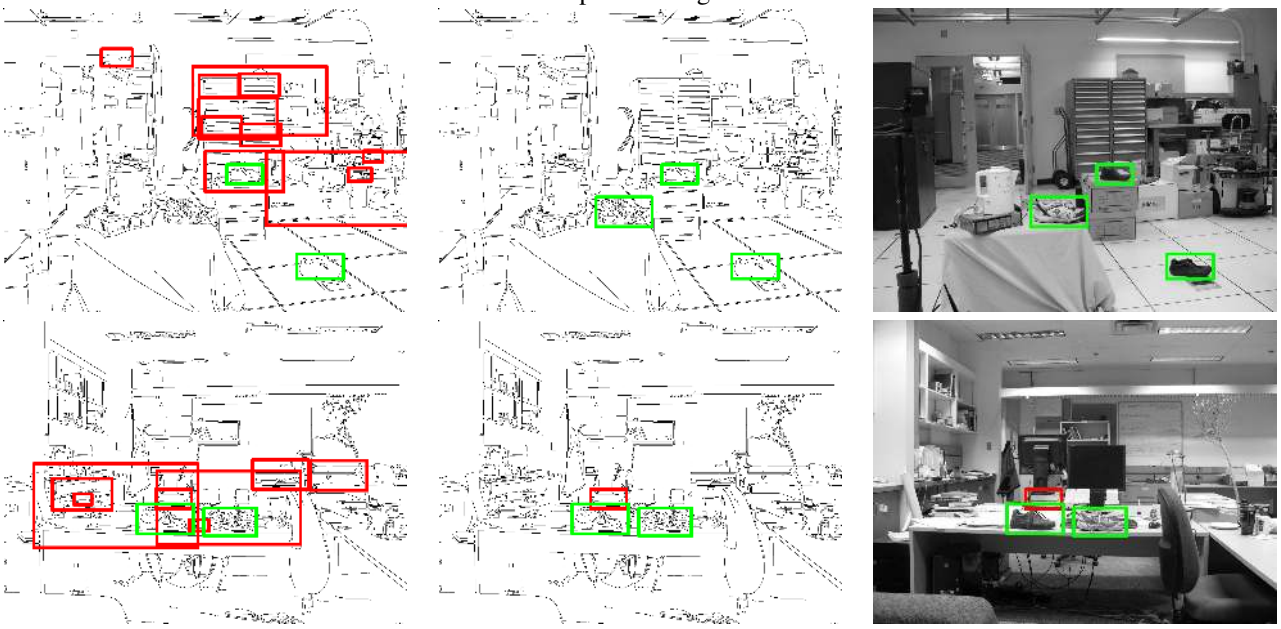
[22] A. Opelt, A. Pinz, and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet," in *CVPR*, 2006.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60(2), pp. 91–110, 2004.

[24] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.



Some examples of mug detections



Some examples of shoe detections

Fig. 5. Examples of object detections. Examples in the left column are from the base classifier at a recall rate at 0.7. Examples in the middle and right columns are from the classifier with the scale prior at the same recall rate. Green signifies true positives and red signifies false positives. As can be seen, the number of false positives is significantly reduced with a scale prior.