Multiple Viewpoint Recognition and Localization

Scott Helmer, David Meger, Marius Muja, James J. Little, David G. Lowe

University of British Columbia

Abstract. This paper presents a novel approach for labeling objects based on multiple spatially-registered images of a scene. We argue that such a multi-view labeling approach is a better fit for applications such as robotics and surveillance than traditional object recognition where only a single image of each scene is available. To encourage further study in the area, we have collected a data set of well-registered imagery for many indoor scenes and have made this data publicly available. Our multi-view labeling approach is capable of improving the results of a wide variety of image-based classifiers, and we demonstrate this by producing scene labelings based on the output of both the Deformable Parts Model of [1] as well as a method for recognizing object contours which is similar to chamfer matching. Our experimental results show that labeling objects based on multiple viewpoints leads to a significant improvement in performance when compared with single image labeling.

1 Introduction

Object recognition is one of the fundamental challenges in Computer Vision. However, the framework in which it is typically evaluated, by labeling bounding boxes within a single image of each scene, is quite different from the scenario present in many applications. Instead, in domains ranging from robotics, to recognition of objects in surveillance videos, to analysis of community photo collections, spatially registered imagery from multiple viewpoints is available. Spatial information can be aggregated across viewpoints in order to label objects in three dimensions, or simply to further verify the uncertain inference performed in each individual image.

This paper proposes such a scene labeling approach, by which we refer to labeling the objects in a scene. We do not choose a particular target application nor tailor the approach to a specific classification function. Instead we present a method that takes multiple well-registered images of a scene and image-space classification results in those images as input and determine an improved set of 3D object locations that are consistent across the images. Our method for locating consistent regions consists of two steps. The first step is a sampling procedure that draws a finite set of candidate 3D locations in order to avoid the high computational cost of considering every potential location. The second step scores these potential locations based on how well they explain the outputs of the image-based classifier in all available viewpoints. Experimental analysis

shows that this method produces significant increases in labeling accuracy when compared against the image-based classifiers upon which it is based.

Figure 1 illustrates a scenario for which a top-performing method on the Pascal Visual Object Categories (VOC) challenge mis-labels several objects in a scene. We have observed that, for such scenes, occlusion and appearance similarity between categories are the most significant challenges for correct recognition. In the left image of Figure 1, two bowls are not detected because their contours are broken by occlusion. Also, the alignment of a bottle and bowl forms a mug-like contour, which causes a false positive for the single-image appearance model in another case. In contrast, labeling from multiple viewpoints achieves correct inference because such accidental alignments occur in only a fraction of the views, and the relatively larger number of views without occlusion support one another to give confident detections. The correct labelings for both scenarios are shown in the right image of Figure 1.

The contribution of this paper is a novel scene labeling strategy based on imagery from multiple viewpoints as well as a new data set suitable for evaluation of such an approach. Our data set contains spatially registered imagery from many viewpoints of a number of realistic indoor scenes. We have made this data set publicly available, as part of the UBC Visual Robot Survey (UBC VRS[1]) as we hope the availability of such data will encourage other authors to consider the problem of recognizing objects from a number of viewpoints, rather than in single still images.

The next section of this paper describes related work in multi-view scene labeling. This is followed by a technical description of our method in Section 3. Next we describe the data set that we have collected, and provide results for our approach evaluated on this data. The paper is concludes with a discussion of future work and outstanding problems.

2 Related Work

View-point independent category recognition is currently an active area of research, with a number of new approaches being advanced, [2–4]. These approaches attempt to perform viewpoint-independent inference, which would, in principle, remove the requirement to have images from multiple viewpoints to annotate a scene. However, these methods typically require annotated training data from a semi-dense sampling of viewing directions and in some cases require additional information such as a video sequence [2]. While viewpoint invariant category recognition is a promising direction, we argue that for certain categories and scenes, multiple viewpoint recognition is advantageous, as in Figure 1.

Integrating information across many images has been a major focus of active vision. Several authors have described Bayesian strategies to combine uncertain information between views, [5, 6]. In particular [6] have previously suggested the use of a generative model of object appearance conditional on the object

[1] http://www.cs.ubc.ca/labs/lci/vrs/index.html

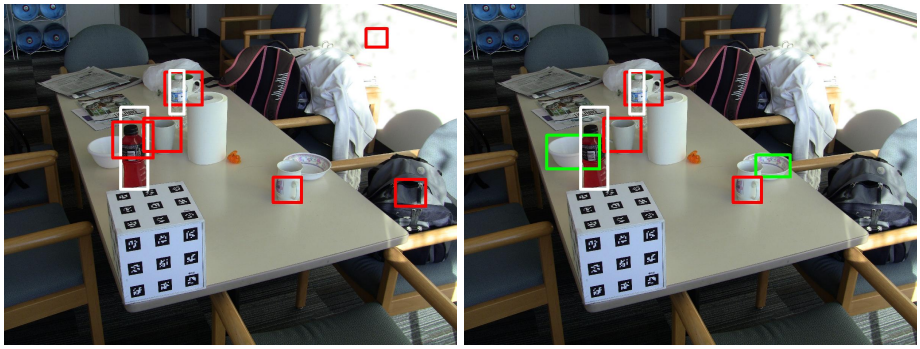


Fig. 1. On the left is an image labeled with the Deformable Parts Model from [1], a state-of-the-art approach to recognize objects in single images. On the right is a result from our multi-view approach. The challenges presented by occlusion and clutter are overcome by fusing information across images. Bowls are labelled in green, mugs in red and bottles in white. A threshold equivalent to 0.65 recall has been used for both methods.

label and other confounding variables such as pose and lighting, along with a sequential update strategy in order to solve this problem. However, active vision has typically been used to recognize object instances or categories for which accurate pose estimation is possible. We extend some of these ideas for more general object categories where accurate pose estimation is still a challenge.

Several authors have also recently considered fusing information across the frames of a video sequence to improve over single-frame performance. For example, Andriluka *et al.* [7] use a bank of viewpoint-dependent human appearance models and combine these with a learned motion prior to gather consistent information across motion tracks. Also, Wojek *et al.* [8] infer the location of pedestrians simultaneously with the motion of a vehicle to achieve localization in 3D from an on-board camera. The probabilistic scoring function described in Section 3 is similar to the approaches used in each of these methods.

The recent work by Coates and Ng [9] most closely resembles our own, although developed independently. Here, they first use multiple images and rough registration information to determine possible corresponding detections, using similar techniques to us. The posterior probability for each set of corresponding detections is computed, where non-maximum suppression is used to discard errant correspondences. Their work differs from ours most significantly in that the posterior probability for a correspondence is based solely on appearance, where as our work includes geometric information as well. In addition, their experimental validation is limited, presenting multi-view results on a single category, where the difference in viewpoint is not particularly significant. Our work presents a formulation that is more general with more extensive experiments to demonstrate the utility of multi-viewpoint recognition.

Numerous robotic systems with object recognition capabilities have also been previously presented. Many systems are targeted for outdoor navigation such as

intelligent driving systems (see several recent vision based methods among far too many others too mention [10, 11]), however in most cases these systems have attempted to recognize just a few of the most relevant object categories for the task of safe navigation: typically pedestrians, cars, in some cases stop signs. In contrast, indoor robots attempting to perform human-centric tasks that require a broader sampling of visual categories include [12–15]. Many of these systems have provided us with inspiration and share aspects of our own approach. However, we are not aware of any platform that reasons so explicitly about the integration of information across viewpoints.

Another contribution of our work is the publication of a new dataset, called the UBC VRS. There are several existing data sets that contain imagery of objects from multiple viewpoints, similar to the one described in this paper. Savarese *et al.* [16] is perhaps the most similar since it focuses on challenging object categories and each image shows an object instance in a realistic scene. However, this data set has only a single instance in each image, the objects occupy nearly all of the image and there is little clutter. In each of these aspects, the data presented in this paper is more challenging. Several other datasets also feature imagery from multiple viewpoints, which is intentionally either less realistic, less cluttered or both [17, 18]. To our knowledge, the data presented in this paper represents the largest available set of *spatially-registered* imagery for realistic scenes that is both publicly available and annotated with the presence of challenging object categories.

3 Method

We define multiple-viewpoint object localization as the task of inferring the 3D location, scale and pose of a set of objects from image-space detections in well-registered views. Each image-space hypothesis represents a ray in 3D, so objects observed from several views with matching detections will produce many nearly intersecting rays. This set of rays should mutually agree upon position, location and scale of a single 3D object. Our method involves locating a set of objects, C, that maximizes the the conditional likelihood of the set of image-space detections, F: $p(F|C)$.

Section 3.1 describes our likelihood function in more detail. Then, Section 3.2 grounds our discussion by describing 2 image-space detectors that we have used as inputs. In Section 3.3, a method for proposing candidate object locations based on single factors within the full model is developed. This technique proposes a larger number of candidate objects than is optimal, and so a final assignment and scoring procedure returns a final set of inferred objects, as is described in Section 3.4.

3.1 Likelihood Model

In order to describe the likelihood model in detail, we expand upon the definitions of 3D objects c and image-space detections f. Each $c \epsilon C$ has a 3D position X, and

azimuth orientation θ. As all of the objects considered have a single up-direction and our registration process allowed us to directly observe the gravity direction, elevation angle is neglected here. Each detection f (also referred to as a feature) consists of a category label, a bounding box b, a response v, and (depending on the detector) an azimuth orientation θ. We seek to optimize $p(F|C)$, and this requires a mapping h, such that $h(c) = F_c$ where F_c is the set of detections accounted for by object c. We assume that every detection is generated by at most one 3D object, and we enforce that all detections not assigned to a cluster as assigned to a null cluster. Briefly, valid assignments are those for which the 3D object projects near to the detected bounding box. We will expand upon our search for h shortly. For now, we express the goal of our inference as maximizing:

$$p(F|C) = \sum_h p(F, h|C) \tag{1}$$

$$= \sum_h p(F|h, C)p(h|C) \tag{2}$$

$$\simeq \sum_h \prod_{c \in C} \prod_{f \in h(c)} p(f|h, c)p(h|C) \tag{3}$$

$$\tag{4}$$

In Equation (4) we assume detections f are conditionally independent given an assignment to generating objects. We approximate the above with q(C),

$$q(C) = \max_h \prod_{c \in C} \prod_{f \in h(c)} p(f|h, c)p(h|C) \tag{5}$$

$$\tag{6}$$

Therefore, the target for maximization is our selection of potential objects in the scene C.

The above approach is not uncommon in object classification. It is similar in spirit to the constellation models of Fergus *et al.* [19], with the exception that our features are detector responses from multiple viewpoints.

We continue to decompose Equation (6) in order to express it in terms of the geometric priors available to our system. Define d_f as the distance from the camera centre to the projection of X onto the Z axis of the camera for which detection f occurred, z_f as the focal length of that camera and X_f as the reprojection of X into the image. Given a mapping, we define the score for an object c as similar to the first term in Equation (6), that is:

$$score(c) = \prod_{f \in h(c)} p(f|h, c) \tag{7}$$

$$= \prod_{f \in h(c)} p(v_f, \theta_f, b|c) \tag{8}$$

$$= \prod_{f \in h(c)} p(v_f|c)p(\theta_f|\theta_c)p(|b_{centre} - X_f||X)p(b_{scale}|cat) \quad (9)$$

The first term in Equation (9) represents the generative appearance model of the detector, discussed above. The second term represents agreement in pose. Here we utilize a Gaussian distribution on the angular difference, with $\mu = 0$ and $\sigma = pi/8$. In the case where the detector does not provide a pose estimate (as is the case with DPM), we omit this term. The third term penalizes distance between the reprojected object centre X_f and the bounding box center for that detection. In this case, $|b_{centre} - X_f|$ is scored using a Gaussian with $\mu = 0$ and $\sigma = \sqrt{b_{area}}/4$, truncated to 0 when X_f lies outside the bounding box. The final term is a scale prior represented as a Gaussian centred about the expected size of each object category. Using z_f, d_f, and the scale prior, the last term is a Gaussian with parameters $\{z_f\mu/d_f, z_f\sigma/d_f\}$.

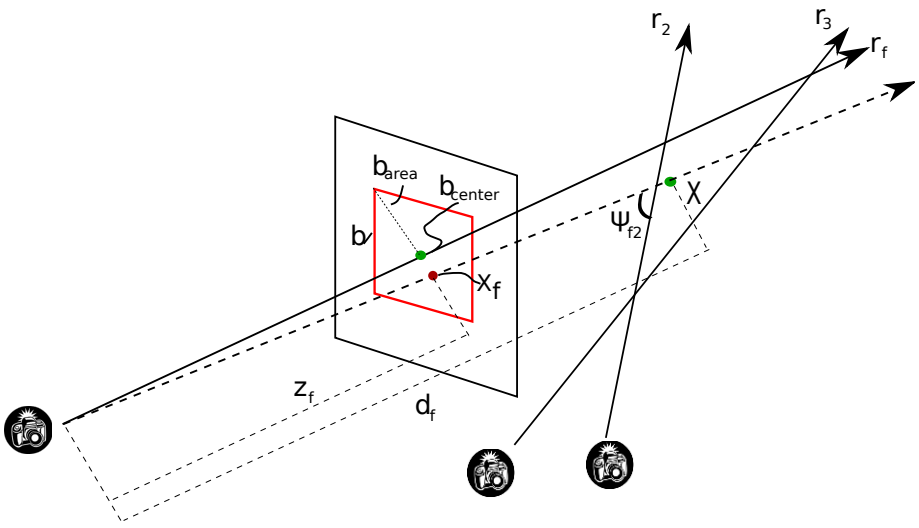


Fig. 2. The geometry of the 3D scene labeling problem. Bounding boxes identified by image-based classifiers project to rays (r_f,r_2, r_3) in 3D. Near intersections of rays suggest an object's 3D location, X, if the suggested scale (using d_f, z_f, b_{center}) agrees with the scale prior and the reprojection of X onto the image plane, X_f, is close to b_{center}. If an azimuth pose is suggested by the detector, then we can utilize ψ_{f2} as well to determine if the detections agree on an object pose θ.

3.2 Object Detectors

Our approach is not dependent upon a particular object recognition technique. Instead, we can produce scene labelings based on any object detector that produces a finite number of responses, f, each detailing a bounding box b, a score v,

and possibly a pose estimate for the object θ. Ideally, we have a generative model for the classifier. That is, we know the probability of each score value v given the presence of the object class. We can utilize validation data to build an empirical distribution for v. In our implementation, we have utilized two different object classifiers to demonstrate the applicability of our approach.

Deformable Parts Model The Discriminatively Trained Deformable Part Model (DPM) [1] was one of the best performing methods in the 2009 Pascal VOC. DPM is based on a complex image feature that extends the Histogram of Oriented Gradients (HOG) descriptor [20]. HOG describes an object as a rigid grid of oriented gradient histograms. The more recent DPM approach extends HOG by breaking the single edge template into a number of parts with probabilistic (deformable) locations relative to the object's centre. DPM detections are weighted based on the margin of the SVM to produce a real-valued score. We translate this score into an observation model, v from above, with a validation step. The classifier is originally trained using a set of hand collected imagery from the internet, along with other publicly available datasets. We have employed DPM classifiers for 3 of our studied object categories: Mugs, Bottles, and Bowls.

Boundary Contour Models We wanted to explore the possibility of using object pose in our method, so we implemented a simple classifier that not only outputs a probability v, but also a pose. We have discretized the azimuth into 8 viewpoints, and represent each viewpoint as a separate classifier. The classifier for one viewpoint has as its model a single boundary contour. For a particular window in the image, the edges are compared to a scaled version of the boundary contour using a version of oriented chamfer matching [21], and this distance is represented as v. Using a validation set we have empirically modeled, $p(v|cat, \theta)$, the distribution of v when the object with pose θ is present in the image. This classifier is used in the sliding window paradigm, using non-maximum suppression to return a finite set, F, of responses $f = (v, b, \theta)$, as required. The training and validation data we used for the shoe classifier came primarily from Savarese *et al.* [16], with a few additional images acquired from internet imagery.

3.3 Identifying Potential 3D Objects

We seek an efficient way to find a set of objects C that will maximize Equation (6). This is accomplished by casting rays passing from the camera's centre through the centre of the bounding box. The size of the bounding box, along with the scale prior, suggests where along the ray that a potential object could be located. With a multitude of rays in the scene, locations of near-intersection for an object category suggest the presence of an object. See Figure 2 for an example.

Determining a reasonable set of near-intersections can be challenging, depending on the nature of the scene and the false positive rate of the object

detector. For all pairs (i, j) of rays for a particular category, we use four properties to construct a boolean adjacency matrix, A, that indicates rays that might correspond to the same 3D object. First, i and j cannot come from the same image. Second, the minimum distance between the rays must satisfy: $d_{i,j} < 0.5\mu$, where μ and σ are the scale priors for the category. Third, the scale of the object, s_i, (suggested by the bounding box size and distance along the ray i) must satisfy: $\|s_i - \mu\| < 2\sigma$. Finally, for a classifier that supplies pose information, the rays must agree somewhat on the azimuth angle of the object. More precisely, the angle between the two rays, ψ, must be within $\pi/4$ of the expected angle between the two detections. We apply these hard thresholds in order to produce a boolean adjacency matrix, and significantly reduce the potential near-intersections that must be considered in later stages.

Using A, for all valid pairs of rays (i, j) we compute the 3D point X that minimizes the reprojection error between X and the bounding box centres in the both image planes. This X becomes a potential object c. Then, we again utilize A to determine potential agreeing rays (i.e. constructing h) to compute $score(c)$, including only those rays for which c explains their geometric quantities (bounding box size, and position) better than a uniform prior. In the case of detections that also return pose, we also infer the best object pose using the rays that are assigned to c. The result of this process is a much larger set of objects than are likely.

3.4 Maximum Likelihood Object Localization

The final step in our approach is to determine a set of candidate objects that approximately optimizes the likelihood function. We use a greedy strategy to select final objects from the large set of candidates proposed previously, and construct a matching h. That is, we score each object c, and iteratively add the one achieving the highest score to C, afterwards assigning its supporting detections, F_c to c, ie $h(c) = F_c$. We then remove these detections and their rays from all remaining potential objects. Following this, we recompute the scores, and repeat this process until all detections f are assigned to an object $c \in C$, or to the null object. We will finally end up with a matching h, the objects C, and a score for each of the objects $c \in C$.

At this stage, we have an assignment of each 2D detection to a 3D potential object in the case that multi-view consensus was found, or to nothing if it was not. We attempt to use this matching to re-score the input image-space detections such that they reflect the inferred geometry as well as possible. If a detection has been mapped to an object, we assign the score of the object to the 2D detection. If the detection has mapped to a null object, the score remains what it would have been in the single view case since we could bring no additional information to explain the geometric quantities.

4 Experiments

We have evaluated the scene labeling performance of our technique using the UBC VRS dataset, a collection of well-registered imagery of numerous real-world scenes. The next Section will describe this data set in detail. Section 4.2 will subsequently describe an experimental technique for fair evaluation of scene labeling approaches. It also contains the results generated from these approaches, which illustrate the performance of our technique.

4.1 Collecting Images from Multiple Registered Viewpoints

Numerous interacting factors affect the performance of a multi-view registration system. Many are scene characteristics, such as the density of present objects, the appearance of each instance and the environment lighting. Others are artifacts of the image collection process, such as the number of images in which each object instance is visible at all and whether its appearance is occluded. Ideally, we would like to evaluate our technique on imagery with similar properties to likely test scenarios. As discussed in Section 2, existing datasets are not suitable to evaluate realistic scene labeling because they either lack significant clutter or are generated synthetically.

Therefore, we have collected a new dataset, the UBC VRS, containing a variety of realistic indoor scenes imaged from a variety of viewpoints. Each scene contained many of our evaluation object categories without our intervention. In a few cases, we have added additional instances in order to increase the volume of evaluation data, but we have been careful to preserve a realistic object distribution. The physical settings present in the dataset include 11 desks, 8 kitchens and 2 lounges. In addition, we have augmented the highly realistic scenes with several "hand-crafted" scenarios, where a larger than usual number of objects were placed in a simple setting. We have assembled 7 shoe-specific, and 1 bottle-specific scene of this nature.

As mentioned, each scene has been imaged from a variety of viewpoints, and each image has been automatically registered into a common coordinate frame using a fiducial target of known geometry. Fiducial markers are a common tool for tasks ranging from motion capture for the movie industry to 3D reconstruction. Our target environment involves highly cluttered, realistic backgrounds, and so simple coloured markers or uniform backgrounds (i.e. green screens) are not desirable. Instead, we have constructed a 3D target from highly unique visual patterns similar to those described in [22–24]. This target can be robustly detected with image processing techniques, and image points corresponding to known 3D positions (marker corners) can be extracted to sub-pixel accuracy. For the experiments in this paper, we have estimated a pinhole model for our cameras offline, so these 2D-3D correspondences allow the 3D pose of the camera to be recovered.

When evaluating practical inference techniques aimed at realistic scenarios, repeatability and control of experiments is of highest importance. In order to allow other researchers to repeat our experiments, we have released the entire

set of imagery used for to generate all of the following results as part of the UBC VRS dataset at the address http://www.cs.ubc.ca/labs/lci/vrs/index.html.

4.2 Evaluation

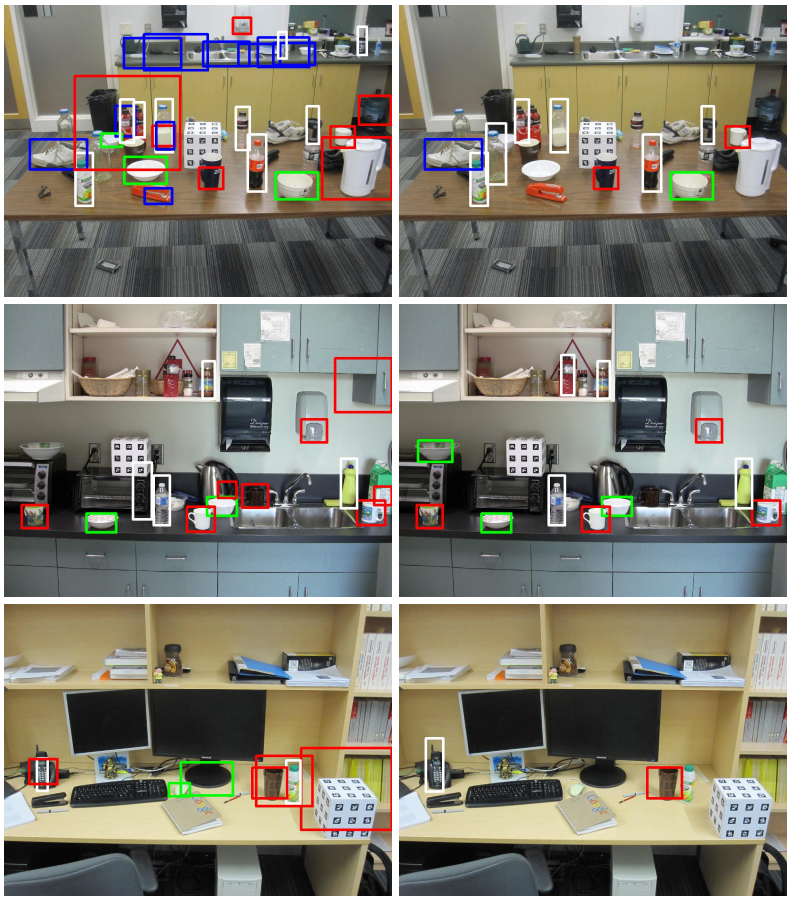
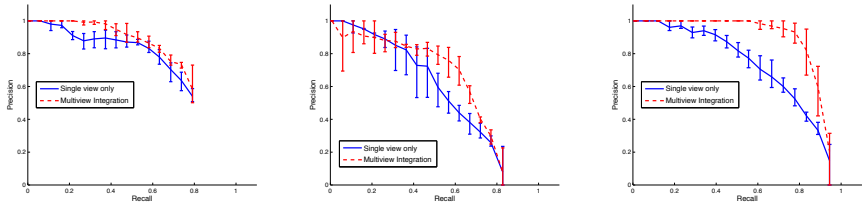


Fig. 3. Image-based detections (left) and multi-viewpoint detections from our method (right). Mugs are shown in red, shoes in blue, bowls in green, and bottles in white. A 0.65 recall threshold of used for all categories but shoes which use recall of 0.25.

To measure localization performance, we compare the output of our automated labeling procedure with ground truth annotations produced by a human labeler. Our labeling procedure follows the the Pascal VOC format, which is a well-accepted current standard in Computer Vision. Specifically, each object is labelled using a tight bounding box and 2 additional boolean flags indicate

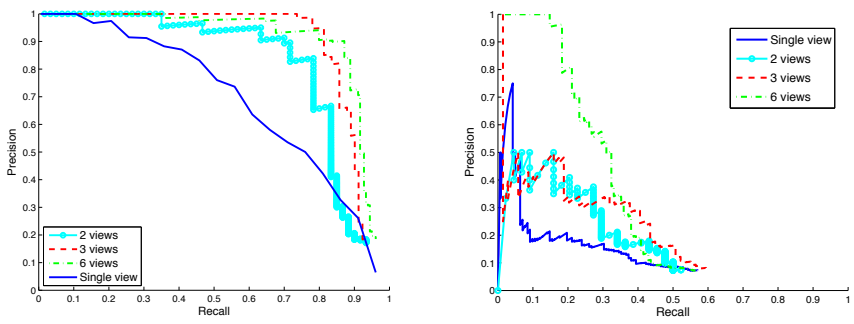


Bottles Mugs Bowls

Fig. 4. The above Recall-Precision curves show the multi-view approach as compared to the single view approach when the number of viewpoints available is fixed to 3.

whether the object is truncated (e.g. due to occlusion) and/or difficult (e.g. visible, but at an extremely small scale). Instances flagged as difficult or truncated are not counted in numerical scoring.

We also employ the evaluation criterion used for the VOC localization task. That is, each object label output by a method is labeled as a true or false positive based the ratio of area of intersection vs area of union between the output bounding box and the closest ground truth annotation of the same category. A precision-recall curve is used to summarize detection results over a variety of possible thresholds, and this curve can be summarized into a single value by summing the area under the curve (AUC).



Bowls Shoes

Fig. 5. The performance of our system generally increases as the number of views for each scene is increased.

Our first experiment utilizes the evaluation criteria described to compare the scene labeling produced by our method with the labeling produced by image-space classification methods. For each scene, we perform numerous trials of labeling, to achieve statistical significance. In each trial we select a sub-set of 3 images obtained from well-separated viewpoints. Trials are made independent by

randomizing the starting location for this viewpoint selection, such that the labeling procedure sees mostly non-overlapping sets of images between trials. The results of all trials over all scenes in the testing database are shown in Figure 4.

The multi-view approach significantly outperforms labeling based on single images. This is somewhat expected given that a multi-view approach can utilize more information. We have analyzed the situations where the multi-view procedure is not able to infer the correct labeling, and comment on these briefly here. First, we note that there are situations where the appearance based detector simply fails, suggesting further work on the object detectors. Second, there are a number of objects that cause inter-category confusion, even a low recall. For example, the top of a mug or a plate look similar to a bowl in most viewpoints. This could be remedied by including structure information or priors that preclude different objects occupying the same space. We leave this for future work.

We have also studied the contribution that several system components make to our labeling accuracy, and here we describe the effect of each. First, we varied the number of viewpoints considered for each scene. For brevity, only the results for the category *bowl* are shown in Figure 5 and the results for the remaining categories are displayed in a more compact form in Table 1. The general trend is that additional viewpoints lead to more accurate labeling. There is however, a notable difference in the behaviour between classes identified with the DPM detector (mug, bowl, bottle) and those identified with the contour detector (shoe).

For the mug, bowl and bottle, the addition of a second view of the scene yields a significant increase in performance, a third view gives strong, but slightly lessening improvement, and further additional views begin to yield less and less improvement. Our analysis of this trend is that the DPM detector gives sufficiently strong single-view performance, that after aggregating information across only a small number of images, nearly all the object instances with reasonably recognizable appearances are identified correctly. Adding additional viewpoints beyond the third does increase the confidence with which these instances are scored, but it can no longer change the labels such that an instance moves from incorrect to correct, and thus only modest improvement in the curve is possible.

On the contrary, the result from the shoe detector is interesting in that the performance for two viewpoints is little better than for a single image, but the performance does increase significantly between the third and sixth image considered. Our analysis of these results shows that this trend results from the relatively lower recall of the shoe detector for single images. In this case considering more viewpoints increases the chance of detecting the shoe in at least a portion of the images. Moreover, since the shoe detector is sensitive to pose, accidental agreement between hypotheses is unlikely.

Finally, we examined the effect of the scale prior on labeling performance. Table 1 demonstrates that the AUC score improves for each of the classes considered when the scale prior is applied. The use of scale improves the set of clusters that are proposed by improving the adjacency matrix, and it also improves the accuracy of MAP inference for cluster scoring.

Number of Views	1	2	3	6
Mugs	0.57	0.60	0.65	0.67
Bottle	0.67	0.75	0.76	0.75
Bowl	0.71	0.79	0.86	0.90
Shoe	0.1	0.13	0.18	0.28

Scale Prior	Disabled	Enabled
Mugs	0.60	0.65
Bottle	0.69	0.76
Bowl	0.84	0.86

Table 1. A summary of results generated when evaluating our approach for a variety of object categories. Each value in the table summarizes precision and recall over all possible thresholds with the area under the curve (AUC).

5 Conclusions

This paper has presented a multi-view scene labeling technique that aggregates information across images in order to produce more accurate labels than the state-of-the-art single-image-classifiers upon which it is based. Labelling scenes from many viewpoints is a natural choice for applications such as the analysis of community photo collections and semantic mapping with a mobile platform. Our method is directly applicable to applications where accurate geometry has been recovered, and as our results demonstrate, the use of information from many views can yield a significant improvement in performance.

References

1. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: In Proceedings of the IEEE CVPR. (2008)
2. Su, H., Sun, M., Fei-Fei, L., , Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: Proceedings of the IEEE International Conference on Computer Vision. (2009)
3. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Gool, L.V.: Using multi-view recognition and meta-data annotation to guide a robot's attention. International Journal of Robotics Research (2009)
4. Liebelt, J., Schmid, C., Schertler, K.: Viewpoint-independent object class detection using 3d feature maps. In: In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2008)
5. Whaite, P., Ferrie, F.: Autonomous exploration: Driven by uncertainty. Technical Report TR-CIM-93-17, McGill U. CIM (1994)
6. Laporte, C., Arbel, T.: Efficient discriminant viewpoint selection for active bayesian recognition. International Journal of Computer Vision **68** (2006) 1573 – 1405
7. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA (2010)
8. Wojek, C., Roth, S., Schindler, K., Schiele, B.: Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes. In: In proceedings of ECCV. (2010)
9. Coates, A., Ng, A.Y.: Multi-camera object detection for robotics. In: IEEE International Conference on Robotics and Automation. (2010)

10. Leibe, B., Schindler, K., Cornelis, N., Gool, L.V.: Coupled object detection and tracking from static cameras and moving vehicles. IEEE Transactions on Pattern Analysis and Machine Intelligence (2008)
11. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR. (2009) 1–8
12. Kragic, D., Björkman, M.: Strategies for object manipulation using foveal and peripheral vision. In: IEEE International Conference on Computer Vision Systems ICVS'06. (2006)
13. Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Meissner, M., Bradski, G., Baumstarck, P., Chung, S., Ng, A.: Peripheral-foveal vision for real-time object recognition and tracking in video. In: Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07). (2007)
14. Rusu, R.B., Holzbach, A., Beetz, M., Bradski, G.: Detecting and segmenting objects for mobile manipulation. In: ICCV, S3DV Workshop. (2009)
15. Ye, Y., Tsotsos, J.K.: Sensor planning for 3d object search. Computer Vision and Image Understanding **73** (1999) 145–168
16. Savarese, S., Fei-Fei, L.: 3d generic object categorization, localization and pose estimation. In: IEEE Intern. Conf. in Computer Vision (ICCV), Brazil (2007)
17. Viksten, F., Forssen, P.E., Johansson, B., Moe, A.: Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In: In proceedings of the IEEE International Conference on Robotics and Automation (ICRA). (2009)
18. LeCun, Y., Huang, F., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2004)
19. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV). (2005)
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2., San Diego, USA (2005) 886 – 893
21. Shotton, J., Blake, A., Cipolla, R.: Multiscale categorical object recognition using contour fragments. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 1270–1281
22. Fiala, M.: Artag, a fiducial marker system using digital techniques. In: CVPR'05. Volume 1. (2005) 590 – 596
23. Poupyrev, I., Kato, H., Billinghurst, M.: Artoolkit user manual, version 2.33. Human Interface Technology Lab, University of Washington (2000)
24. Sattar, J., Bourque, E., Giguere, P., Dudek, G.: Fourier tags: Smoothly degradable fiducial markers for use in human-robot interaction. In: Fourth Canadian Conference on Computer and Robot Vision (CRV), Montreal, Quebec, Canada (2007) 165–174