Detecting Dynamic Objects with Multi-View Background Subtraction

Raúl Díaz*, Sam Hallman*, Charless C. Fowlkes
Computer Science Department, University of California, Irvine
{rdiazgar,shallman,fowlkes}@ics.uci.edu

Abstract

The confluence of robust algorithms for structure from motion along with high-coverage mapping and imaging of the world around us suggests that it will soon be feasible to accurately estimate camera pose for a large class photographs taken in outdoor, urban environments. In this paper, we investigate how such information can be used to improve the detection of dynamic objects such as pedestrians and cars. First, we show that when rough camera location is known, we can utilize detectors that have been trained with a scene-specific background model in order to improve detection accuracy. Second, when precise camera pose is available, dense matching to a database of existing images using multi-view stereo provides a way to eliminate static backgrounds such as building facades, akin to background-subtraction often used in video analysis. We evaluate these ideas using a dataset of tourist photos with estimated camera pose. For template-based pedestrian detection, we achieve a 50 percent boost in average precision over baseline.

1. Introduction

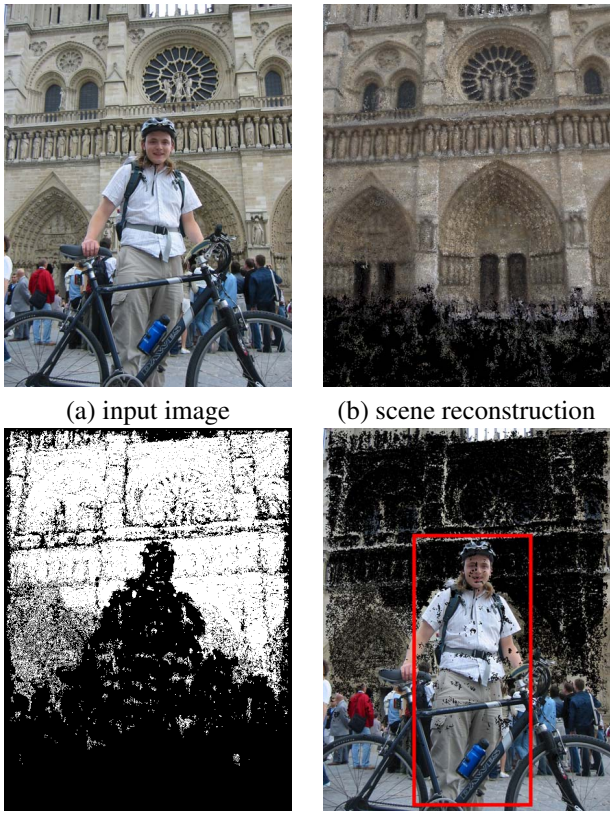
Consider an image of a popular tourist destination shown in Figure 1. How can we exploit the large set of photographs available online depicting this same general location in order to better understand the content of this particular image? It is useful to divide scene components into two categories: **dynamic objects** such as people, bikes, cars, pigeons or street vendors that move about and are likely to only appear in a single image taken at a particular time and **static backgrounds** such as buildings, streets, landscaping, or benches that are visible in many images taken in the same general location.

For static (rigid) backgrounds, a classic approach to scene understanding is to use structure-from-motion (SfM) and multi-view stereo (MVS) techniques to build up an explicit model of the scene geometry and appearance. Such a model can make strong predictions about a novel test image including the camera pose and locations of scene points within the image. These methods are now well developed and work robustly on large unstructured photo collections [22, 7]. For dynamic objects, the problem is less constrained. However, past images of a scene can still provide general information about where objects are likely to appear in the future. For example, we might expect *a priori* to see pedestrians on a sidewalk and cars in the middle of the street (and not vice versa). This idea has been explored extensively in the literature on scene context [25, 13] and more recently in work on affordances [11, 10].

While images of real scenes typically contain both static and dynamic components, these corresponding approaches to scene understanding have largely been pursued independently. Work on scene context the last few years has focused on single-image geometry estimation (e.g. [17, 14, 11]) since stereo or other depth estimates were often unavailable. On the other hand, from the perspective of multi-view geometry, dynamic objects are a nuisance and must be treated as outliers during matching. Here we explore how to combine these two ideas, namely: *How can strong models of static backgrounds improve detection of dynamic objects?*

We propose two different approaches that utilize static scene analysis for detection. The first is to perform unsupervised analysis of a large set of scene images in order to automatically train **scene-specific object detectors**. At test time, if we have rough camera localization (e.g., GPS coordinates), we can invoke the appropriate scene-specific detector rather than a generic detector. It seems obvious that an object detector trained with data from a specific scene has the potential to perform better than a generic detector since it can focus on modeling specific aspects of a scene which may be discriminative. If resources are available to perform ground-truth labeling for images collected from every possible scene location, we could simply use existing methods to train a large collection of specialized detectors (one for each object category appearing in each possible scene). However, this is not a scalable solution as it requires labeling positive examples in each possible scene as well as training a huge bank of object detectors. Our key observa-

* authors contributed equally to this work



(a) input image (b) scene reconstruction

(c) background mask (d) detected foreground

Figure 1: Wide-baseline matching to a collection of photos provides estimates of which pixels belong to static background regions. (a) shows an input image and (b) shows the re-projection of a 3D model built from other images of the same scene. While this model is not realistic enough to allow for direct comparison of pixel values, because we have 3D structure we can easily compare the appearance of local image patches in the input image to corresponding patches in the image set used to build the model. (c) shows those patches for which this match score was above a given threshold. (d) based on this parsing of the scene into static background and dynamic foreground objects, we can eliminating spurious false-positives and improve object detector performance.

tion is that while acquiring scene-specific positive training instances is expensive, it is possible to automatically produce large quantities of scene-specific negative training instances in an unsupervised manner by identifying portions of a scene that are likely to be static background.

The second approach which we term **multi-view background subtraction** is inspired by a classic trick used to analyze video surveillance data or webcam image streams. When a scene is repeatedly imaged by a fixed camera, one can build up a model of the scene background (e.g. by computing the median color of a pixel over a sequence of images) and compare it to a new image (subtraction) in order to segment out regions that are likely to correspond to dynamic objects of interest. Unfortunately, such a model is tied to the pixel coordinate system and hence offers little help for understanding a new image taken from a novel viewpoint or with a different camera. If instead we model the static background in world coordinates (e.g., as a high-quality 3D mesh) and accurately estimate the camera pose for a test image, we can render the appropriate background image and perform subtraction as before to identify static and dynamic image regions. While this might have seemed like an unrealistic idea even ten years ago, the availability of robust algorithms for SfM and MVS along with high-coverage mapping and imaging of the world suggests that high-quality 3D models and precise camera localization of novel photos will be soon be commonplace for a variety of scenes, particularly urban outdoor environments (see, e.g., [19]).

At their core, both of these approaches tackle the same problem of modeling static background for a scene. Scene-specific object detectors implicitly contain a model of the scene background derived from negative training examples. Since the detectors are used in a sliding window fashion, this model of the background is translation invariant and must function well at any image location. Multi-view background subtraction goes one step further by synthesizing a spatially varying model of the background. The detector then competes with the background model in order to explain the image contents at each image location. A key distinction is that the former works during training to generate a large collection of object detectors while the later demands more substantial test-time inference. In our experiments, we find both approaches useful and often provide independent benefits in detection performance.

In the remainder of the paper we discuss the SfM and MVS tools we use to analyze image collections, give specifics of the scene-specific background model and multi-view background subtraction approaches, and finally describe a set of experiments evaluating their efficacy. We conclude with a brief discussion of related work.

2. Isolating Backgrounds with Multi-View Stereo

We propose to use large photo collections in a unsupervised manner to build up a model of the static rigid background appearance in a given scene. Such photos will necessarily contain non-static objects but these can generally be rejected as they are not consistent from one photo to the next. Our basic technical tools are robust structure-from-motion and multi-view stereo.

2.1. Recovering Camera Pose and Scene Geometry

There is a large body of work which has emerged over the last few years on the problem of camera localization and large scale SfM [22, 12, 1, 5]. We use an off-the-shelf software pipeline to reconstruct the static scene in which our objects are placed. After computing SIFT descriptors for a collection of image keypoints [16], we use Bundler [21] which performs sparse keypoint matching and bundle-adjustment in order to estimate scene structure and camera pose from a large collection of un-calibrated images. Once camera poses have been estimated, we use PMVS [8] to perform dense reconstruction using multi-view stereo. Since stereo matches do not need to be computed across all pairs of views, we use CMVS [7] in order to perform view clustering prior to running PMVS which significantly increases the speed of reconstruction. This generally yields high quality reconstructions like that one shown in Figure1(b).

When presented with a novel test image, we would like to similarly estimate the camera calibration and pose. This can again be accomplished using SIFT keypoint matching to find correspondences and standard methods for camera calibration from epipolar geometry. Since our dataset is small we use simple batch processing with Bundler. In a real system, such matching can be carried out incrementally with high accuracy and accelerated with fast indexing in order to scale perform matching to large world-wide datasets [19, 15]. For example [15] demonstrate rapid indexing millions of images and tens of millions of keypoints.

2.2. Identifying Background Pixels

Given a high-quality 3D model of a scene and a known camera pose and calibration, it is straightforward to synthesize an image from that viewpoint as shown in Figure1(b). Comparison of this re-projected scene with the actual image should indicate which pixels that differ from the static scene and hence are likely to be dynamic objects of interest. Unfortunately, simply computing the difference between the re-projected image and the test image does not work well in practice. While the models generated by the above pipeline are quite compelling, they are not pixel-perfect. Renderings of point cloud models typically lack fine-scale features that provide cues for object detection. Furthermore, because the library of images used in constructing the model are taken across a huge range of lighting conditions with different cameras, even when the recovered geometry is perfect, the estimated average color may be quite different than the particular color that appears in a novel test image.

While one could develop a image differencing scheme that is robust to these variations, we observe that the problem of *determining when an image from a novel viewpoint matches a model is exactly the problem that multi-view stereo algorithms are designed to solve!* Rather than comparing an image patch to the model, we compare it directly to the appearance of corresponding patches in the images in our test library. We describe the basic matching function we use and refer the reader to [8] for more details.

Consider a point p on our scene reconstruction which is predicted to be visible in our test image I. Let $V(p)$ be the set of all images in our image collection that depict this same point (including only those views where the point p is visible based on the reconstruction). We compute a measure of photometric discrepancy between the image collection and our test image given by

$$match(p) = \frac{1}{|V(p)|} \sum_{J \in V(p)} h(p, I, J) \qquad (1)$$

where $h(p, I, J)$ compares the color at a set of points sampled from a local plane tangent to the static background reconstruction at p and projected into the test image I and each other image J using the recovered camera poses. These sample points lie on a 5x5 grid on the tangent plane. Their color is estimated from each image using bilinear interpolation and the colors compared using normalized correlation.

Using this match score we generate a **background score map** that indicates the quality of match for each pixel to the images in the dataset. Where appropriate, we can threshold this score map $match(p) > \alpha$ to yield a binary **background mask** as shown in Figure1(c). In our experiments we used a threshold correlation score of $\alpha = 0.5$ but the system performance is very robust to this choice (see supplement). To compute patch matches in our test and training we used a modified version of the publicly available PMVS software [8, 6] which implements the necessary patch matching functionality in order to compute background masks. Note that while we use the same discrepancy function as PMVS, our goal is slightly different. Stereo reconstruction only requires finding a few high-scoring matches across the whole image set in order to estimate geometry and color of the point p. Once it has found enough views, it is free to ignore many images in which p may actually be visible. In our case, we would like to estimate a dense collection of match scores over the entire surface visible from the test image even if this particular test image does not offer the best match for the point. Additionally, our choice of maximal discrepancy threshold used in producing the mask is higher than would typically be used in matching in order to provide denser support in each individual image.

In the following two sections, we describe two different ways in which to use this background mask. First, at training time to generate negative training examples in an unsupervised manner. Second, at test time to prune detection responses which fire on regions that are estimated to be background.

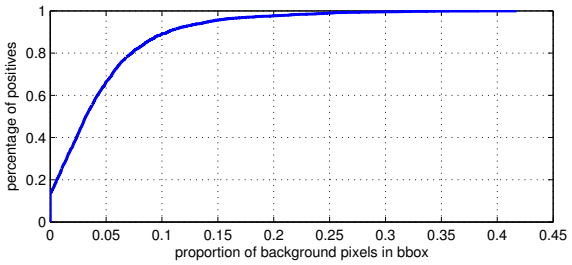


Figure 2: Cumulative distribution of the proportion of background pixels $q(i)$ inside true-positive object instances in the scene specific training set.

3. Training Detectors with Scene Specific Background Models

A standard approach to detection is to train a sliding window classifier that distinguishes the object of interest from background. We propose to use the information about the scene derived automatically from a collection images of that scene in order to tune the detector to perform better in that particular context. This can be accomplished by selecting negative training instances from those regions of the image that are expected to be background based on the match score (Equation 1).

Rather than including all possible negative windows of an image, we utilize a standard approach of hard-negative mining in order to generate a concise collection of negative instances with which to train the detector. Given an initial estimate of the template (e.g., derived from a generic training set), we run the detector on images (or parts of images) known not to contain the object. Any location where the detector responds at a level greater than the SVM margin specified by the current weight vector is added to the pool of negatives as it may constitute a support vector. This process of hard-negative mining and retraining of the classifier are interleaved until no further negatives are found at which point the final weight vector is computed.

When ground-truth annotations of positives for a scene specific dataset are available, hard negative mining can easily be used by just dropping any candidate negative windows that overlap significantly with a ground-truth positive. However, labeling images is a labor intensive process. Instead we use the background mask as a proxy that can be produced in an unsupervised manner. Let B_i be the set of pixels inside a bounding box associated with a candidate detection i. We compute the proportion of background pixels in this region as:

$$q_i = \frac{1}{|B_i|} \sum_{p \in B_i} (match(p) > \alpha)$$

Figure 2 shows the distribution of the background mask proportions, q_i, over the set of true-positive detections in our training dataset. We use a conservative criteria, declaring a candidate window background if $q_i > 0.2$. As can be seen in the figure, by and large positive instances are not confused with static background. Using this criteria, less than 3% of the ground-truth positives are incorrectly judged as part of the background by this criteria.

4. Multi-View Background Subtraction

Background-subtraction has long been used in the surveillance community as it provides a useful approach to scene segmentation (see, e.g., [23]). A closely related problem is that of video stabilization which yields background subtraction in the case of video with relatively high frame rates [20]. Our scenario differs in that the images we consider may be taken from very different cameras, sparsely sampled in time, different lighting, etc., which make tracking-based approaches used in video inappropriate. Instead we use SfM to estimate the camera parameters of a novel test image and then utilize the same technique described in Section 2 for identifying background pixels, namely those that are photo-consistent with our model and image collection.

For a novel image, we can view the background mask as a hypothesized segmentation and ask if the detection is consistent with this segmentation. Motivated by previous work on combining segmentation and detection, we investigated several mechanisms for measuring consistency with the background mask. This included examining consistency with average shape masks derived from example segmentations of each object and explicitly learning a mask template from example training data (see Experiments). We also tested using GrabCut [18] or super-pixels in order to refine the background mask estimate based on local image evidence such as discontinuities in color and texture. In the end we found that simply using the proportion of background mask pixels inside the bounding box with the same simple threshold used in generating scene specific negatives ($q(i) > 0.2$) was as effective at removing false-positives as any of these more elaborate schemes.

5. Experimental Results

In order to evaluate performance of our ideas we constructed a labeled dataset of pedestrians from a collection of images of Notre Dame provided online by the authors of Bundler [21]. Images were annotated with bounding boxes around each pedestrian using a protocol similar to the PASCAL VOC detection set. Bounding boxes were tight around the visible portion of the person. Difficult examples such as seated or heavily occluded people were included in the ground truth but tagged as "difficult" and not used in computing the test performance numbers presented here (i.e., they neither count as true or false positives). Low resolu-

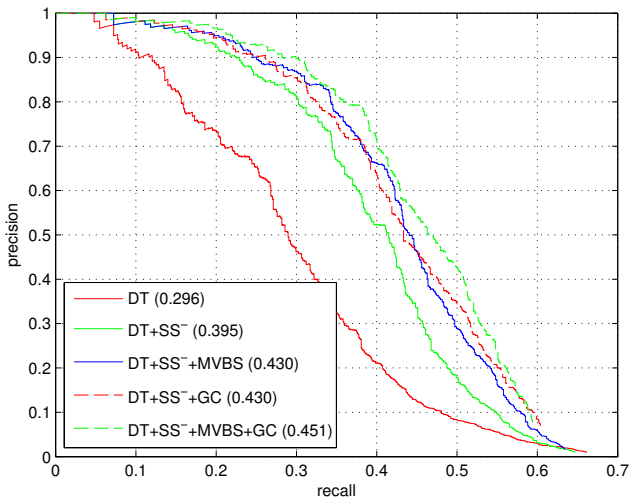


Figure 3: Precision-Recall for pedestrian detection with scene specific detectors. **DT** is the baseline Dalal-Triggs template detector trained on the INRIA dataset. **+SS$^-$** is trained using scene-specific negative instances mined in an unsupervised manner from images of Notre Dame. **+GC** prunes detections where the bottom of the detection appears above the horizon based on the camera pose estimated using SfM. **+MVBS** prunes detections whose bounding box contain more than 20% estimated background pixels based on multi-view matching. The scene-specific model performs significantly better than the baseline with multi-view background-subtraction and geometric consistency both providing additional gains in detector precision.

	DT	DT+SS$^-$	DPM	DPM+SS$^-$
Detection	0.296	0.395	0.455	0.551
+MVBS	0.412	0.430	0.558	0.552
PoP [13]	0.323	0.322	0.348	0.323
PoP+SfM	0.405	0.406	0.404	0.337

	DT+FS$^-$	DT+FS	DPM+FS$^-$	DPM+FS
Detection	0.41	0.43	0.55	0.63

Figure 4: Average Precision for pedestrian detection with scene-specific detectors. **DT** is the baseline Dalal-Triggs template detector [2] and **DPM** is the deformable parts model of [4]. **+SS$^-$** indicates the detector was trained with automatically acquired scene-specific negative instances. **+MVBS** prunes detections whose bounding box contain more than 20% estimated background pixels based on wide-baseline matching. **PoP** shows the results obtained by using the approach in [13], while **+SfM** enforces the horizon estimation using structure from motion results. The unsupervised scene-specific model performs significantly better than the baseline, with multi-view background-subtraction providing additional gains in detector precision. For comparison we also show performance for fully supervised scene-specific training. **+FS** indicates results using scene specific positive and negative instances, **+FS$^-$** uses only negtive instances. Our unsupervised approach achieves similar levels of performance and is scalable to large numbers of scenes.

tion people (< 40 pixels) were marked as difficult or not annotated. Benchmarking used the standard PASCAL detection benchmark criteria in which 50% overlap between a detection and ground-truth bounding box is sufficient (where overlap is the ratio of intersection area to area of union).

The 401 images available were randomly split into 200 test and 201 training. The training images were used when automatically generating negative examples to train the scene-specific detector as well as during algorithm development to validate the choice of bounding box mask threshold parameter and SVM regularization.

Baseline Detectors: We focus our experiments on two popular object detectors. First, we consider an implementation of the Dalal-Triggs (DT) rigid template model [2]. We train our implementation of the detector on positive and negative examples provided in the INRIA Person dataset. The resulting baseline detector achieves an AP=0.79 on the INRIA test set, comparable to the results reported elsewhere [2]. Performance of this baseline detector on the Notre Dame test dataset (AP=0.296) is lower than on INRIA due to the greater variety of appearances of pedestrians labeled in the ND dataset (e.g. more poses, occlusion and truncation, wider range of scale, etc.) We set the SVM regularization parameter C to maximize performance on the Notre Dame training images.

In addition to the DT model, we also test with the *deformable parts model* (DPM) of Felzenswalb *et al*. [4], using the implementation from [9] trained on the PASCAL VOC 2007 (train+val) dataset [3]. This model is substantially more complex but achieves better baseline performance (AP=0.455) by modeling deformation as well as mixture components which better detect truncated people marked in the dataset.

Training Scene-Specific Background Models: The 201 training images were used in building the scene specific background model (denoted **DT+SS$^-$** and **DPM+SS$^-$** in the figures). For this purpose we did *not* use the ground-truth annotations but did utilize the background mask with the $q_i > 0.2$ background threshold. We found that both models performed significantly better when trained with scene specific negatives. DT improved from 0.296 to 0.395 and DPM from 0.455 to 0.551 average precision (compare columns of Figure 4). We did find it necessary to decrease the degree of regularization when changing the size of the training data set (C went from 0.1 to 0.01).

We also compared our scene-specific background model with unsupervised hard-negative mining to a supervised version (**FS-**) in which the scene-specific negatives were chosen to not overlap with any positive bounding boxes by more than 10%. This achieved an AP of 0.41 for DT and 0.55 for DPM suggesting that our unsupervised negative mining based on masks is capturing most of the useful negative examples. Finally, we evaluated fully supervised versions (**FS**) of the DT and DPM models which include both scene-specific positives and negatives in addition to the INRIA and PASCAL training sets, respectively. This yielded an AP of 0.43 for DT and 0.63 for DPM respectively.

In training the scene specific model, it is useful to start with a pretrained model and then perform additional passes of hard negative mining on the scene specific images. We found that this hot-starting was significantly more efficient than retraining the model from scratch. For example, training the DT detector from scratch took 83 minutes compared to only 37 minutes to hot start. Similarly, the DPM model takes days to train on the whole PASCAL dataset but only hours to hot start.

Multi-View Background Subtraction: We tested the multi-view background subtraction scheme (MVBS) using simple thresholding by rejecting detections with $q(i) > 0.2$. Results are shown in Figure 3 and 4. Rejecting such false positives increased the average precision of both detectors. In the case of the DT detector, the combination of MVBS and SS$^-$ training achieves even better performance while the DPM model saturates at 0.55 average precision.

In addition to the simple mask thresholding scheme, we also experimented with learning various features derived from the mask including the mean count of background pixels in the bounding box, the mean match score, and a spatial mask template with various spatial binnings. We found that none of these gave huge performance gains over the simplest thresholding. Learning a spatial mask template on the ND training set with spatial binning at the same resolution of the HOG descriptor gave $AP = 0.476$ while using pixel-sized bins yielded performance of $AP = 0.480$. However, the resulting templates had relatively little structure and are likely over-fit to the statistics of the background masks recovered for this particular scene rather than being universally applicable for all pedestrians.

Figure 5 shows qualitative example outputs of the baseline detector, scene-specific detector and the effect of multi-view background subtraction. There are many textured regions on the cathedral facade where the baseline detector produces false positives. In particular, the carved human figures on the facade naturally match the template well. The model trained with additional scene-specific negatives is able to reject some of the false-positives as it finds very similar examples in the training set which are used as negative support vectors.

Geometric Context: A skeptical reviewer might be concerned that all we are doing is removing those detections up "in the sky", something that could be accomplished using SfM alone without constructing a dense background mask. To check this, we estimated the position of the horizon line based on the recovered camera pose for each test image. Since the plaza in front of the Cathedral is largely planar, we do not expect any pedestrians to appear floating above the horizon. This simple check of geometric consistency also achieves substantial performance improvements for the Dalal-Triggs detector, raising the average precision to $AP = 0.42$. However, multi-view background subtraction is able to prune additional detections which satisfy geometric constraints but include patches of facade visible in the training set providing small but distinct gains over purely geometric pruning (see +GC curves in Figure 3 and qualitative examples in Figure 5). This distinction would be even more obvious in a more complicated scene with elevated structures (balconies, stairs, playground equipment, trolly platforms, etc.)

Putting objects in perspective: We evaluated the Putting Objects in Perspective (PoP) system of Hoiem *et al*. [13] which performs more sophisticated joint probabilistic inference over the camera pose, scene geometry, and detection hypotheses. We considered two different scenarios. In the first we simply substituted our baseline detector but used the camera pose and geometry priors graciously provided by the authors. In the second scenario, we replaced the default prior Gaussian distribution over horizon line position with a tightly peaked double-exponential (b=0.005) centered at the horizon estimate based on SfM camera pose estimation. We also tried using a prior derived from the camera heights produced during bundle adjustment but were unable to find a scaling that yielded better results than the prior from the original paper.

For the default camera pose priors, the PoP inference routine is able to boost the DT detector performance from 0.296 to 0.323. Substituting in the much stronger horizon estimate produced by SfM provides a much more significant boost, up to an average precision of 0.4. Surprisingly, these gains are not present when using the DPM detector. We believe this might be because the conversion of the detector score into a probability based on logistic fitting produces an overestimate of the detector confidence which skews the inference result. We include example detections and horizon estimates produced by PoP in the supplementary material.

6. Discussion

Much of the work on general-purpose object recognition has focused on detecting objects against arbitrary backgrounds of non-objects or other object categories. Such systems are typically trained with negative examples taken

from random images off the web. The idea proposed here is in some ways counter to much contemporary research in category-level object recognition which has focused on generic detectors that will work in a wide range of environments. Indeed, it is a topic of hot debate whether our models are over-fitting to even the most general detection benchmarks[24]. Of course, such over-fitting is useful if you know which dataset (in this case scene) that you will be tested on.

Perhaps the most closely related work to ours is the "Putting objects in perspective" by Hoiem et. al. [13] which makes joint inferences about scene geometry, camera pose and detection likelihoods. PoP only attempts to encode generic prior knowledge about the scene geometry and camera pose in the form of a surface orientation classifier [14]. In contrast, we argue that for many scenes, it is not unreasonable to expect that other photos of the same scene are available from which to do more aggressive geometric reasoning. It thus seems worthwhile to revisit the idea of geometric context in the setting of large-scale SfM which can provide much more reliable estimates of scene geometry for many parts of a novel test image as well as camera pose. From a research perspective, this would help isolate the benefits of geometric context for detection from the difficulties of single-image geometry estimation. Our experiment with pruning detections based on the horizon line from camera pose estimates touches on this but one could clearly go much further. For example, one could utilize the surface estimates returned from multi-view stereo or even re-project a 3D map which was annotated with "affordances" indicating what spatial volumes are likely to contain which objects and in which poses.

Finally, it seems valuable to think how recent work on robust reconstruction relates to problems of recognition. Efforts such as Google Street View that are assembling ever larger collections of images and other data into rich maps and models of city-scapes must constantly deal with dynamic objects (tourists, cars, pigeons, trash, etc.) which constitute outliers to be ignored during matching or better yet eliminated. However, from a broader perspective of scene understanding, *one model's outlier is another model's signal* and these annoyances should be transmuted into useful cue for recognizing dynamic objects.

Acknowledgments

This work was supported by NSF IIS-1253538 and a Balsells Fellowship to RD.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building Rome in a day. *ICCV*, pages 72–79, Sept. 2009. 3

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I: 886–893, 2005. 5

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. Pascal VOC2007 results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–45, Sept. 2010. 5

[5] J. Frahm, P. Fite-Georgel, and D. Gallup. Building Rome on a cloudless day. In *ECCV*, 2010. 3

[6] Y. Furukawa. CMVS. http://grail.cs.washington.edu/software/cmvs/. 3

[7] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Towards Internet-scale multi-view stereo. In *CVPR*, 2010. 1, 3

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *TPAMI*, 1(1):1–14, 2010. 3

[9] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/. 5

[10] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair? In *CVPR*, 2011. 1

[11] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 1

[12] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 3

[13] D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *CVPR*, 2:2137–2144, 2006. 1, 5, 6, 7

[14] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. *ICCV*, 1:654–661, 2005. 1, 7

[15] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide Pose Estimation using 3D Point Clouds. In *ECCV*, 2012. 3

[16] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, Nov. 2004. 3

[17] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. *NIPS*, 2003. 1

[18] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3):309–314, 2004. 4

[19] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. *ICCV*, 2011. 2, 3

[20] Y. Sheikh, O. Javed, and T. Kanade. Background Subtraction for Freely Moving Cameras. *ICCV*, pages 1219–1225, Sept. 2009. 4

[21] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 2006. 3, 4

[22] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz. Scene Reconstruction and Visualization From Community Photo Collections. *Proceedings of the IEEE*, 98(8):1370–1390, Aug. 2010. 1, 3

[23] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *CVPR*, 1999. 4

[24] A. Torralba and A. Efros. Unbiased look at dataset bias. *CVPR*, pages 1521–1528, 2011. 7

[25] A. Torralba and P. Sinha. Statistical context priming for object detection. *ICCV*, 1:763–770, 2001. 1

Figure 5: Example detector outputs at 50% recall. Unsupervised scene specific training makes the detector better able to reject common distractors (e.g. the statues in row 2). MVBS can prune additional false positives at test-time by performing stereo matching to a database of existing images. Note that MVBS is able to remove some false positives which are not caught by geometric consistency (GC) with the horizon line because the hypothesized detections overlap heavily with regions identified as background.