Multi-View Background Subtraction for Object Detection

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

Consider 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 an image taken at a particular time, and **static backgrounds** such as buildings, streets, landscaping, or benches that are visible in many different images taken in the same 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 [8, 1]. For dynamic objects, past images of a scene can 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. This idea has been explored extensively in the literature on scene context [9, 5] and more recently in work on affordances [4, 3].

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. [7, 6, 4]). 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 approaches that utilize static scene analysis for detection. The first is to perform unsupervised analysis of a large set of scene images 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. Our key observation 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.

(a) input image (b) reconstruction (c) bg mask

Figure 1: Wide-baseline matching to a collection of internet photos provides estimates of which pixels belong to static background regions.

The second approach, which we term **multi-view background subtraction**, is inspired by the classic trick used to analyze video surveillance data where one can build up a model of the scene background (e.g. median color) and compare it to a new image (subtraction). 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.

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 ex-

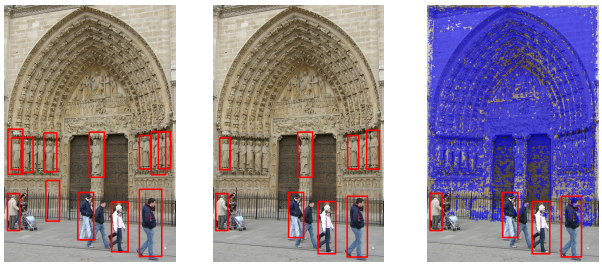
plain the image contents at each image location. A key distinction is that the former works during training to generate an specific object detector for each scene while the later demands more substantial test-time inference.

Experiments: We experiment using an off-the-shelf software pipeline to reconstruct the static scene in which our objects are placed. We use Bundler [8] to perform keypoint matching and bundle-adjustment across SIFT descriptors computed on a large collection of un-calibrated images (a set of 400 images of Notre Dame annotated with pedestrian bounding boxes). Once camera poses have been estimated, we use a modified version of PMVS [2] to perform multi-view stereo yielding a dense set of corresponding patches. We threshold the patch match quality scores to yield a binary **background mask** as shown in Figure1.

Scene Specific Background Models: 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 and can't be carried out for every possible scene. Instead we use the background mask as a proxy that can be produced in an unsupervised manner, dropping candidate scene-specific negatives with less than 80% background points. We found that both the Dalal-Triggs (DT) and Deformable Part Model (DPM) detectors performed significantly better when trained with scene specific negatives (SS^-). DT improved from 0.30 to 0.40 and DPM from 0.46 to 0.55 average precision. Using fully supervised scene-specific negatives yielded 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.

Multi-View Background Subtraction: For a novel image, we ask if a detection is consistent with the segmentation specified by the background mask. We explored a variety of approaches to integrating this information (GrabCut, super-pixels, shape priors, etc.) but found that simply dropping windows with a large proportion of background pixels performed as well. This multi-view background subtraction scheme (MVBS) yielded gains for both detectors (AP of 0.41 and 0.56 for DT+MVBS and DPM+MVBS respectively). In the case of the DT detector, the combination of MVBS and SS^- training achieves even better performance while the DPM model saturates.

The figure above shows the baseline DT detector running at 50% recall, the 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 (e.g., carved human figures 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.



DT DT+SS^- DT+SS^-+MVBS

Geometric Context: We also compared the Putting Objects in Perspective work of Hoiem *et al*. [5] using the recovered camera pose to impose a tight prior on the horizon line position. This yielded AP of 0.40 for DPM, suggesting multi-view background subtraction is able to prune additional detections which cannot be easily pruned from geometric considerations alone. This distinction would be even more obvious in complicated environments (balconies, stairs, playground equipment, trolly platforms, etc.) where the universal ground-plane assumption is violated.

It summary, it 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. From a research perspective, this would help isolate the benefits of geometric context for detection from the difficulties of single-image geometry estimation. The detection model here is simple but 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.

References

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

[2] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *TPAMI*, 1(1):1–14, 2010. 2

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

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

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

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

[7] 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

[8] 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, 2

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