3D Model based Object Class Detection in An Arbitrary View

Pingkun Yan, Saad M. Khan, Mubarak Shah
School of Electrical Engineering and Computer Science
University of Central Florida
http://www.eecs.ucf.edu/~vision

Abstract

In this paper, a novel object class detection method based on 3D object modeling is presented. Instead of using a complicated mechanism for relating multiple 2D training views, the proposed method establishes spatial connections between these views by mapping them directly to the surface of 3D model. The 3D shape of an object is reconstructed by using a homographic framework from a set of model views around the object and is represented by a volume consisting of binary slices. Features are computed in each 2D model view and mapped to the 3D shape model using the same homographic framework. To generalize the model for object class detection, features from supplemental views are also considered. A codebook is constructed from all of these features and then a 3D feature model is built. Given a 2D test image, correspondences between the 3D feature model and the testing view are identified by matching the detected features. Based on the 3D locations of the corresponding features, several hypotheses of viewing planes can be made. The one with the highest confidence is then used to detect the object using feature location matching. Performance of the proposed method has been evaluated by using the PASCAL VOC challenge dataset and promising results are demonstrated.

1. Introduction

In recent years, the problem of object detection has received considerable attention from both the computer vision and machine learning communities. The key challenge of this problem is the ability to recognize any member in a category of objects in spite of wide variations in visual appearance due to geometrical transformations, change in viewpoint, or illumination.

In this paper, a novel *3D feature model* based object class detection method is proposed to deal with these challenges. The objective of this work is to detect the object given an arbitrary 2D view using a general 3D feature model of the class. In our work, the objects can be arbitrarily transformed (with translation and rotation), and the viewing position and orientation of the camera is arbitrary as well. In addition, camera parameters are assumed to be unknown.

Object detection in such a setting has been considered a very challenging problem due to various difficulties of geometrically modeling relevant 3D object shapes and the effects of perspective projection. In this paper, we exploit a recently proposed 3D reconstruction method using homographic framework for 3D object shape reconstruction. Given a set of 2D images of an object taken from different viewpoints around the object with unknown camera parameters, which are called *model views*, the 3D shape of this specific object can be reconstructed using the homographic framework proposed in [10]. In our work, 3D shape is represented by a volume consisting of binary slices with 1 denoting the object and 0 for background. By using this method, we can not only reconstruct 3D shapes for the objects to be detected, but also have access to the homographies between the 2D views and the 3D models, which are then used to build the 3D feature model for object class detection.

In the feature modeling phase of the proposed method, SIFT features [12] are computed for each of the 2D model views and mapped to the surface of the 3D model. Since it is difficult to accurately relate 2D coordinates to a 3D model by projecting the 3D model to a 2D view (with unknown camera parameters), we propose to use a homography transformation based algorithm. Since the homographies have been obtained during the 3D shape reconstruction process, the projection of a 3D model can be easily computed by integrating the transformations of slices from the model to a particular view, as opposed to directly projecting the entire model by estimation of the projection matrix. To generalize the model for object class detection, images of other objects of the class are used as *supplemental views*. Features from these views are mapped to the 3D model in the same way as for those model views. A codebook is constructed from all of these features and then a 3D feature model is built. The 3D feature model thus combines the 3D shape information and appearance features for robust object class detection.

Given a new 2D test image, correspondences between the 3D feature model and this testing view are identified by matching feature. Based on the 3D locations of the corresponding features, several hypotheses of viewing planes can be made. For each hypothesis, the feature points are projected to the viewing plane and aligned with the features in the 2D testing view. A confidence is assigned to each hypothesis and the one with the highest confidence is then used to produce the object detection result.

The rest of the paper is organized as follows. Section 2 provides a brief review of related work. A summary of the 3D shape model construction method is presented in Section 3. The 3D feature model and our object class detection approach are presented in Sections 4 and 5, respectively. Section 6 provides the detection results, and finally, Section 7 concludes the work.

2. Related Work

As the approaches for recognizing an object class from some particular viewpoint or detecting a specific object from an arbitrary view are advancing toward maturity [3, 9, 11], solutions to the problem of object class detection using multiple views are still relatively far behind. Object detection can be considered even more difficult than classification, since it is expected to provide accurate location and size of the object.

Researchers in computer vision have studied the problem of multi-view object class detection resulting successful approaches following two major directions. One path attempts to use increasing number of local features by applying multiple feature detectors simultaneously [1, 6, 13, 14, 15]. It has been shown that the recognition performance can be benefited by providing more feature support. However, the spatial connections of the features in each view and/or between different views have not been pursued in these works. These connections can be crucial in object class detection tasks. Recently, much attention has been drawn to the second direction related to multiple views for object class detection [5, 7, 8]. The early methods apply several single-view detectors independently and combine their responses via some arbitration logic. Features are shared among the different single-view detectors to limit the computational overload. Most recently, Thomas *et al*. [16] developed a single integrated multi-view detector that accumulates evidence from different training views. Their work combines a multi-view specific object recognition system [9], and the Implicit Shape Model for object class detection [11], where single-view codebooks are strongly connected by the exchange of information via sophisticated activation links between each other.

In this paper, we present a unified method to relate multiple 2D views based on 3D object modeling. The main contribution of our work is an efficient object detection system capable of recognizing and localizing objects from the same class under different viewing conditions. Thus, 3D locations of the features are considered during detection and better accuracy is obtained.

3. Construction of 3D Shape Model

A recently proposed homographic framework was employed in our work to construct 3D models from multiple 2D views [10]. A summary of the 3D reconstruction algorithm is provided as follows.

3.1. View Warping and Intersection

Let I_i denote the foreground likelihood map (where each pixel value is the likelihood of that pixel being a foreground) in the ith view of total M views. Considering a reference plane, π_r, in the scene with homography $H_{\pi_r,i}$ from the ith view to π_r, warping I_i to π_r gives the warped foreground likelihood map:

$$\hat{I}_{i,r} = [H_{\pi_r,i}]I_i. \tag{1}$$

The visual hull intersection on π_r (AND-fusion of the shadow regions) is achieved by multiplying these warped foreground likelihood maps:

$$\theta_r = \prod_{i=1}^{M} \hat{I}_{i,r}, \tag{2}$$

where θ_r is the grid of the object occupancy likelihoods plane π_r. Each value in θ_r gives the likelihood of this grid location being inside the body of the object, indeed, representing a slice of the object cut out by plane π_r. It should be noted that due to the multiplication step in (2), the locations outside the visual hull intersection region will be penalized, thus, having a much lower occupancy likelihood.

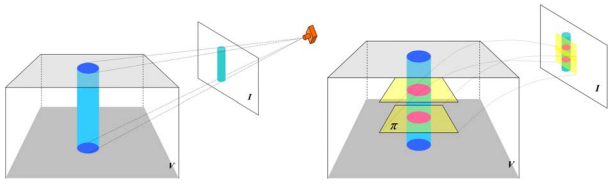
3.2. Construction in Parallel Planes

The grid of the object occupancy likelihood can be computed at an arbitrary number of planes in the scene with different heights, each giving a slice of the object. Naturally this does not apply to planes that do not pass through the object's body, since visual hull intersection on these planes will be empty, therefore a separate check is not necessary.

Let $\mathbf{v}_x$, $\mathbf{v}_y$, and $\mathbf{v}_z$ denote the vanishing points for the X, Y, and Z directions, respectively, and $\mathbf{l}$ be the normalized vanishing line of reference plane in the XYZ coordinate space. The reference plane to the image view homography can be represented as

$$\hat{H}_{ref} = [\mathbf{v}_x \quad \mathbf{v}_y \quad \mathbf{l}]. \tag{3}$$

Supposing that another plane π has a translation of z along the reference direction Z from the reference plane, it is easy



(a) Volume projection (b) Plane transformation

Figure 1. Illustration of equivalence of 3D to 2D projection and plane transformation using homographies. (a) A 2D view of a 3D volume V is generated by projecting the volume on a image plane. (b) The same view can be obtained by integrating the transformation of each slice in the volume to the image plane using homographies.

to show that the homography of plane π to the image view can be computed by

$$\hat{H}_\pi = [\mathbf{v}_x \ \mathbf{v}_y \ \alpha z \mathbf{v}_z + \mathbf{l}] = \hat{H}_{ref} + [\mathbf{I} | \alpha z \mathbf{v}_z] \,, \quad (4)$$

where α is a scaling factor. The image to plane homography H_π is obtained by inverting $\hat{H}_\pi$.

Starting with a reference plane in the scene (typically the ground plane), visual hull intersection is performed on successively parallel planes in the up direction along the body of the object. The occupancy grids θ_i are stacked up to create a three dimensional data structure $\Theta = [\theta_1; \theta_2; \ldots \theta_M]$. Θ represents a discrete sampling of a continuous occupancy space encapsulating the object shape. Object structure is then segmented out from Θ by dividing the space into the object and background regions using the geodesic active contour method [2]. By using the above homography based framework, 3D models for different objects can be constructed.

4. 3D Feature Model Description and Training

In the proposed method, not only the 3D shape of the target object is exploited, but also the appearance features. We relate the features with the 3D model to construct a *feature model* for object class detection.

4.1. Attaching 2D Features to 3D Model

The features used in our work are computed using the SIFT feature detector [12]. Feature vectors are computed for all of the training images. In order to efficiently relate the features computed from different views and different objects, all the detected features are attached to the 3D surface of the previously built model. By using the 3D feature model, we avoid storing all the 2D training views, thus there is no need to build complicated connections between the views. The spatial relationship between the feature points from different views are readily available, which can be easily retrieved when matched feature points are found.

The features computed in 2D images are attached to the 3D model by using the novel homographic framework. Instead of directly finding the 3D location of each 2D feature, we map the 3D points from the model's surface to the 2D views, and find the corresponding features. Our method does not require the estimation of a projection matrix from 3D model to a 2D image plane, which is a non-trivial problem. In our work, the problem is successfully solved by transforming the model to various image planes using homography. Since the homographies between the model and the image planes have already been obtained during the construction of the 3D model, we are able to map the 3D points to 2D planes using homography transformation.

In our work, a 3D shape is represented by a binary volume V, which consists of K slices S_j, $j \in [1, K]$. As shown in Fig. 1(b), each slice of the object is transformed to a 2D image plane by using the corresponding homography $\hat{H}$ in (4). The transformed slice accounts for a small patch of the object projection. Integrating all these K patches together, the whole projection of 3D object in the 2D image plane can be produced. In this way, we obtain the model projection by using a series of simple homography transformations and the hard problem of estimating the projection matrix of a 3D model to a 2D view is avoided.

In our method, the 3D shapes are represented using binary volumes with a stack of slices along the reference direction. Thus, the surface points can be easily obtained by applying edge detection techniques. After transforming the surface points to 2D planes, feature vectors computed in 2D can be related to the 3D points according to their locations. That is the way a 3D feature model is built.

4.2. Beyond the Model Views

The training images in our work come from two sources. One set of images is taken around a specific object of the target class to reconstruct it in 3D as shown in Fig. 2. These images are called *model views*, which provide multiple views of the object but are limited to the specific object. To generalize the model for recognizing other objects in the same class, another set of training images is obtained by using Google image search. Images of objects in the same class with different appearances and postures are selected. These images are denoted as the *supplemental views*.

Since the homographies between the supplemental images and the 3D model are unknown, features computed from the supplemental images cannot be directly attached to the feature model. Instead, we utilize the model views as bridges to connect the supplemental images to the model as illustrated in Fig. 2. For each supplemental image, the model view, which has the most similar viewpoint is specified. The supplemental images are deformed to their specified view by using an affine transformation alignment. Then we can assume that each supplemental image will have the

Figure 2. Construction of 3D feature model for motorbikes. 3D shape model of motorbike (at center) is constructed using the model views (images on the inner circle) taken around the object from different viewpoints. Supplemental images (outer circle) of different motorbikes are obtained by using Google's image search. The supplemental images are aligned with the model views for feature mapping. Feature vectors are computed from all the training images and then attached to the 3D model surface by using the homography transformation.

same homography as the model's corresponding view. The 2D features computed from all of the supplemental training images can now be correctly attached to the 3D model surface using the same method as discussed for the model views. A codebook is constructed by combining all the mapped features with their 3D locations.

5. Object Class Detection

Given a new test image, our objective is to detect objects belonging to the same class in this image by using the learnt 3D feature model M. Each entry of M consists of a code and its 3D locations $\{c, l_c^3\}$. Let F denote the SIFT features computed from the input image, which is composed by the feature descriptor and its 2D location in the image $\{f, l_f^2\}$. Object O_n is detected by matching the features F to the 3D feature model M.

In our work, feature matching is achieved in three phases. In the first phase, we match the features by comparing all the input features to the codebook entries in Euclidean space. However, not all the matched codebook entries in 3D are visible at the same time from a particular viewpoint. So, in the second phase, matched codes in 3D are projected to viewing planes and hypotheses of viewpoints are made by selecting viewing planes with the largest number of visible points projected. In the third phase, for each hypothesis, the projected points are compared to 2D matched feature points using both feature descriptors and locations. This is done by iteratively estimating the affine transformation between the feature point sets and removing the outliers with large distance between corresponding points. Outliers belonging to the background can be rejected during this matching process. The object location and bounding box is then determined according to the 2D locations of the final matched feature points. The confidence of detection is given by the degree of match.

6. Experimental Results

The proposed method has been tested on two object classes: motorbikes and horses. For the motorbikes, we took 23 model views around a motorbike and obtained 45

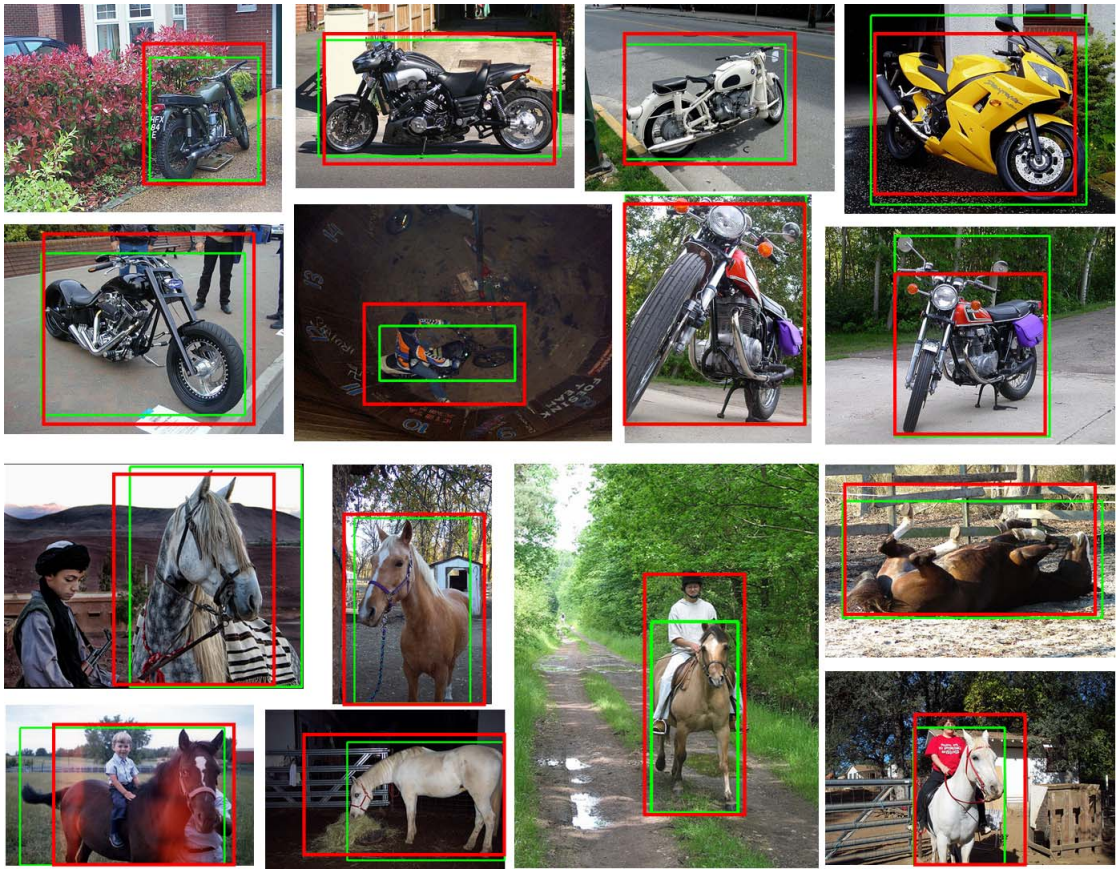


Figure 3. Detection of motorbikes and horses using the proposed approach. The ground truth is shown in green and red boxes display our detected results.

supplemental views by using Google's image search. Some training images of the motorbikes and the 3D shape model are shown in Fig. 2. For the horses, 18 model views were taken and 51 supplemental views were obtained.

To measure the performance of our 3D feature model based object class detection technique, we have evaluated the method on the PASCAL VOC Challenge 2006 test dataset [4], which has become a standard testing dataset for objective evaluation of object classification and detection algorithms. The dataset is very challenging due to the large variability in the scale and poses, the extensive clutter, and poor imaging conditions. Some successful detection results are shown in Fig. 3. The green box indicates the ground truth, while our results are shown in red boxes.

For quantitative evaluation, we adopt the same evaluation criteria used in PASCAL VOC challenge, so that our results can be directly comparable with [8, 16, 4]. By using this criteria, a detection is considered correct, if the area of overlap between the predicted bounding box B_p and ground truth bounding box B_{gt} exceeds 50% using the formula

$$\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} > 0.5. \qquad (5)$$

The *average precision* (AP) and *precision-recall* (PR) curve can then be computed for performance evaluation.

Fig. 4(a) shows the PR curves of our approach and the methods in [8, 16] for motorbike detection. The curve of our approach shows a substantial improvement over the precision compared to the method in [8], which is also indicated by the AP value (0.182). Although our performance is lower than that of [16], considering the smaller training image set used in our experiments, this can be regarded as satisfactory. Fig. 4(b) shows the performance curves for horse detection. While there is no result reported in the VOC challenge using researchers' own training dataset for this task, we compared our result to those using the provided training dataset. Our approach performs better than the reported methods and obtained AP value of 0.144. It is noted that the absolute performance level is lower than that of motorbike detection, which might be caused by the non-rigid body deformation of horses.

7. Conclusion

In this paper, we have proposed a novel multi-view generic object class detection method based on 3D object

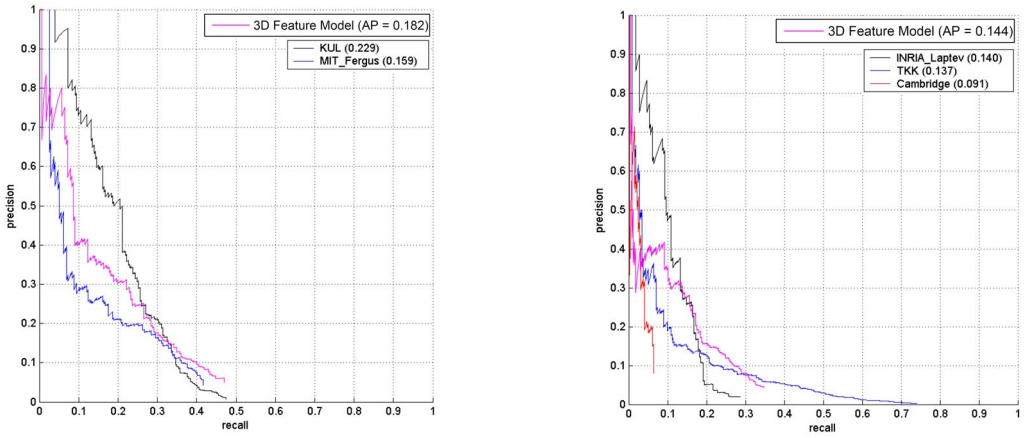


Figure 4. The PR curves for (a) motorbike detection and (b) horse detection using our 3D feature model based approach. The curves reported in [4] on the same test dataset are also included for comparison.

shape and appearance modeling. The proposed method builds a 3D feature model for establishing spatial connections between different image views by mapping appearance features to the surfaces of a 3D shape. Through a novel homographic framework, our method exploits the relationship between multiple 2D views in a more unified manner. Experimental evaluation of the proposed method suggests collaborative information in the 2D training images can be represented in a more unified way through a 3D feature model of the object. We have also revealed that both appearance and shape can be salient properties to assist in object detection. Performance of the proposed method has been evaluated using the PASCAL VOC challenge dataset and promising results have been demonstrated. In our future work, we plan to extend our method by taking supplemental views in a more automated fashion. So, more supplemental views can be easily incorporated to improve the performance. We will also test the method on objects classes with more complex shapes and appearances.

Acknowledgments

This research was funded in part by the U.S. Government VACE program.

References

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR (1)*, pages 26–33, 2005. 2

[2] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997. 3

[3] H. Chang and D.-Y. Yeung. Graph laplacian kernels for object classification from a single example. In *CVPR (2)*, pages 2011–2016, 2006. 2

[4] M. Everingham, A. Zisserman, C. Williams, and L. van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf. 5, 6

[5] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2005. 2

[6] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV*, pages 1134–1141, 2003. 2

[7] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005. 2

[8] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, pages 443–461, 2005. 2, 5

[9] V. Ferrari, T. Tuytelaars, and L. Van Gool. Integrating multiple model views for object recognition. In *CVPR*, volume 2, pages 105–112, 2004. 2

[10] S. M. Khan, P. Yan, and M. Shah. A homographic framework for the fusion of multi-view silhouettes. In *ICCV*, 2007. 1, 2

[11] B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *DAGM*, pages 145–153, 2004. 2

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, Nov. 2004. 1, 3

[13] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV (4)*, pages 490–503, 2006. 2

[14] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *ICCV*, pages 503–510, 2005. 2

[15] E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages 1331–1338, 2005. 2

[16] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, volume 2, pages 1589–1596, 2006. 2, 5