Unfolding Warping For Object Recognition

Jun Xie

Janelia Farm Research Campus, Howard Hughes Medical Institute xiej@janelia.hhmi.org

Min Hu, Mubarak Shah School of Electrical Engineering & Computer Science, University of Central Florida

Abstract

In practice, understanding the spatial relationships between the surfaces of an object, can significantly improve the performance of object recognition systems. In this paper we propose a novel framework to recognize objects in pictures taken from arbitrary viewpoints. The idea is to maintain the frontal views of the major faces of objects in a global flat map. Then an unfolding warping technique is used to change the pose of the query object in the test view so that all visible surfaces of the object can be observed from a frontal viewpoint, improving the handling of serious occlusions and large viewpoint changes. We demonstrate the effectiveness of our approach through analysis of recognition trials of complex objects with comparison to popular methods.

1. Introduction

Object recognition from 2D images is a heavelyresearched problem in the study of both human and machine vision. One of the challenges for automatic object recognition systems is that the same object may appear differently, depending on viewpoints, lighting conditions, and other environmental factors. A possible solution is to match a test image against an explicitly represented 3D geometric model [4]. While explicit models provide powerful geometric constraints, generating the explicit geometric model is typically a difficult and time-consuming process.

An alternative way to solve the above problem is to employ the appearance-based object recognition methods which are intuitive and robust. Rothganger *et al.* [7] imposed the multiple view geometry constraint on potential matching patches and stitched matches found in successive images into a global 3D affine model of the object. Selinger *et al.* [6] proposed assembling the local features within a loose global context and combining the evidences from different viewpoints based on the feature frequency distribution over the entire database in a Bayesian framework. Ferrari *et al.* [2] used a similar intermediate grouping scheme between the primitives and views.

Despite impressive results, the multi-view approaches have some limitations which restrict their application in realistic scenes. Firstly, those approaches usually require storing a large number of views for each object [8] or limiting the range of admissible viewpoints [1]. Moreover, they need accurate viewpoint and pose detection for the query object in the test image. Unfortunately, detecting a large number of correct correspondences is one of the most difficult problems in object recognition. Although there have been some efforts (e.g. [7]) to improve the detection of initial correspondences, correct viewpoint detection is still limited by the number of model views and the object's pose in the test image. In addition, since most model features may appear multiple times in the training dataset, the inherent redundancy makes the multiview recognition extremely inefficient.

This paper presents a novel framework to recognize objects in images taken from arbitrary viewpoints. The basic idea is to represent the model within a global compact view, called a flat map, so that the query object can be compared with the model from the frontal views. In this way, the projective distortions of its side faces can be reduced, resulting in better detection of promising features (see Fig. 3). Finally, we present a recognition system which can handle objects composed of planar surfaces such as boxes and cars.

2. Object Modeling Using Global Flat Map

One way to generate a global flat view for a 3D object is to use the Map Projection (MP) approach, which allows one to represent curved surface on a flat surface (*e.g.*the earth map). However, MP is not adequate for recognition due to the introduced spatial distortion in the resulting map. If the object has a cylindrical or conic shape, the unfolded map of the object's surface can be constructed through the context mosaicing approach [3].

In our case, we take several model views for each model object. These model views can be divided into two classes. The first class consists of the snapshots of the major faces of the object. These frontal views accurately capture the details of the object, providing a strong confidence for object recognition from multiple viewpoints. Another class consists of 'bridge' views in which two or more faces can be simultaneously observed. The model views can be captured individually or extracted from a sequence around the object. If the initial view taken for a certain face has considerable projective distortion, it can be resolved by employing the image rectification technique (*e.g.* [9]).

To combine those views into a compact map, one can sequentially register the multiple-views, which may result in large registration errors, because the registration error might be propagated when multiple pairwise estimates of registration transformation are concatenated. In order to suppress the registration error for the whole object, we choose the star network topology to construct the flat map. More specially, for each bridge image all visible faces are detected by registering them with the front views. Then, the detected frontal views are assembled and the region shared by two connected views is stored in an index table according to the spatial ordering provided by the bridge views. To make the resulting map uniform in intensity, a color rectification process is employed before the assembling process.

3. Registration Using Unfolding Warping

When performing recognition, the input is a single image of an unknown object. It may be a frontal view of one of the object's faces or may contain multiple faces with occlusion and deformation. Given the test image, we extract the Scale Invariant Feature Transform (SIFT) features [5] in both the test image and the model map. Feature correspondences are then identified using a fast nearest-neighbor algorithm. In practice, the initial matching may contain significant errors, which may severely disturb the estimation of homography and is especially serious for side faces due to the large distortions. In addition, if there is a large difference between the test image and the model views, the matches may typically be found on only a small portion of the object. Densely covering the visible part of the object is desirable because it increases the evidence for its presence, providing a higher discriminative power.

Here we propose a scheme for simultaneously increasing both the accuracy of homography estimation and the number of confident correspondences. In particular, for a test image I_t and the model map I_m , we apply the RANSAC scheme to find the meaningful homographies from the initial correspondences. Notice that different homographies may have a different number of inliers. Let $\{\mathbf{H}_y\}|_{y=1}^{Y}$ be the automatically detected homographies and \mathbf{s}_y be the inliers of \mathbf{H}_y . We call the homography with the maximal inliers the dominant homography \mathbf{H}_d .

In order to refine the estimation, we change the



Figure 1. Unfolding warping for object recognition.

viewpoint for each detected face by applying the initial homographies to the test image so more features on those side faces can be observed and matched. Instead of applying all the homographies to the test image, we choose only the dominant one, such that $\hat{I}_t = \mathbf{H}_d I_t$, because this dominant homography is induced by the most significant face visible in the test view. Since the exact segmentation of the face region is difficult, we warp the whole test image instead of the region covered by the corresponding face.

After the homography transformation the warped image is expected to be more similar, in the dominant face area, to the model map than the original one. We then re-compute the features of the warped image, \hat{I}_t , and explore the correspondences with the model map again. These two steps, homography estimation and image warping, iterate until the inliers don't change anymore.

Figure 2 shows an example of this homography refinement process. Initially, the dominant homography has 166 inliers. After the first refinement, the number increases to 844, which is a significant improvement. Figure 2(c) shows the warped image after applying the final dominant homography to the original test image. The resulting homography induces the dominant plane visible in the image pair. Before refining other homographies, we remove the features on the detected face area by re-projecting the inliers $\hat{\mathbf{s}}_d^t \in \widehat{I}_t$ to the original test image: $\mathbf{s}_d^t = \mathbf{H}_d^{-1} \hat{\mathbf{s}}_d^t$. The features on the original test image, except for the re-projected features s_d^t , are used to find subsequent homographies. This process is repeated until the number of remaining correspondences is too small, lower than a predetermined threshold. In this way, the mismatches from side faces to the dominant face can be avoided efficiently. The last row in Fig. 2 shows a side face and its inliers detected automatically using this scheme.

4. Object Recognition

In order to obtain correct recognition based on the features which are generally detected locally, we propose incorporating the spatial context information into the recognition scheme. Consider a set of correspondences $P = \{p_i\}|_{i=1}^n$ and $Q = \{q_i\}|_{i=1}^n$ on the warped test image \hat{I}_t and the model map I_m , respectively. For feature point $p_i \in \hat{I}_t$, its context feature is defined as a



Figure 2. Dominant homography refitment: (a) the model flat map, (b) the test image, and (c) the initial matching using SIFT features (289 pts). Second row shows (d) the initial inliers (166 pts) for the fist dominant face, (e) the inliers (844 pts) of the final dominant homography, and (f) the warped test image with the final homography. Last row: the number of inliers increases from 10 (g) to 134 (h) after two iterations and the warped image (i).

log-polar histogram ψ_i^t , with respect to other features in \hat{I}_t :

$$\psi_i^t(b) = \#\{p_j \neq p_i : (p_j - p_i) \in bin(b)\}, \ p_j \in P$$
(1)

where $b = \{1 \dots B\}$ denotes the index of bins used to describe the feature points. The resulting histogram records the relative locations of other feature points to the central point p_i . Concretely, we use five bins for distance measure $\log r$ and 12 bins for angle θ . Thus, there are a total of 60 bins in the log-polar space, B = 60.

With this descriptor, the spatial context distance of the matches p_i and q_i can be estimated by computing the difference of their context histograms, (*i.e.* the number of features in each sector areas around the focused features). However, since the features are found in different faces, they have different reliability in describing the occurrence of p_i and q_i . On the other hand, we notice that the intensity profile along the line (path) going between two feature points may change greatly if the features lie on different faces and those faces are configured in different ways. The votes from different features should therefore be treated differently according to the profiles of their paths to the focus feature point.

In our method, the voting weight of a feature pair (p_k, q_k) to the center features (p_i, q_i) is determined by the profile difference between the two paths $\overline{p_i p_k}$ and



Figure 3. Test images in the box dataset.

Table 1. Recognition rates (%) of two methods for different test datasets: T_1 (only bridge images) and T_2 (with unseen images).

Dataset	# images	SIFT Matches	Our Method
T_1	124	62.1	78.9
T_2	56	38.63	59.9
T_1+T_2	180	54.8	73.1

 $\overline{q_i q_k}$:

$$\omega_i^k = \sqrt{\frac{[\hat{I}_t(\overline{p_i p_k}) - I_m(\overline{q_i q_k})]^2}{D_{ik}}}, \quad k = 1, 2, \dots, n,$$
(2)

where $I(\overline{a_1 a_2})$ refers to the image intensity profile along the path from point a_1 to point a_2 in image I, and D_{ik} is the distance between (p_i, q_i) and (p_k, q_k) . Then the confidence for each bin in the context grid is defined as:

$$\mathcal{W}_i(b) = \frac{\sum_{p_s \in \text{bin}(b)} \omega_i^s}{\psi_i^t(b)} + \frac{\sum_{q_s \in \text{bin}(b)} \omega_i^s}{\psi_i^m(b)}.$$
 (3)

Now, consider the context around the feature pair. The overall matching cost of feature pair (p_i, q_i) is computed by comparing their context descriptors as:

$$C_i^{tm} = C(p_i, q_i) = \frac{1}{2} \sum_{b=1}^{\mathcal{B}} \frac{\mathcal{W}_i(b) \cdot [\psi_i^t(b) - \psi_i^m(b)]^2}{\psi_i^t(b) + \psi_i^m(b)}.$$
(4)

Once the cost C_i is obtained from each correspondence, the final step is to perform the classification by selecting the model which is the shortest distance from the test image: $\mathcal{D}_{tm} = 1/n \sum_{i=1}^{n} C_i^{tm}$.

5. Experimental Results

We are reporting results of two classes: boxes and cars. For the boxes, the training dataset consists of 21 different boxes, each box having 9 model views. The test dataset contained 180 images, including the bridge images from the training dataset and new ones with occlusions and viewpoint changes as shown in Fig. 3. The SIFT features were extracted from both the model views and the test images, and an initial matching was established. Based on the quantity and quality of those SIFT matches, a classification process was applied for a basic recognition. The results are shown in Table 1.

For our algorithm, we first generated the model flat maps of boxes. Then during testing, we warped the test



Figure 4. Recognition results for boxes. Blue circle, red star and green cross indicate correct, error and non-class result respectively.

image using each model map and determined matches between the model and test image. The model flat map which gave the shortest distance was selected as the match. The recognition results are shown in Fig. 4, where the blue circles represent the correct classifications and the red stars indicate errors. Since the final correspondences in our approach are strictly qualified by geometrical constraint (homography), there may be some test images with no sufficient and accurate matches for meaningful homography estimation. Those cases are put into the non-class category and are indicated with the green cross in Fig. 4. The statistics in Table 1 show that our scheme performs very well in this box database and gives a significant improvement over the SIFT-based matches.

Another dataset tested in our experiments consisted of 17 car objects. Figure 5 shows some of the model flat maps generated for the those cars. The test dataset contained 205 images taken under different situations. Note that most of the cars have two side faces that were similar, or exactly the same in appearance. This is a serious problem for appearance-based matching techniques because a given feature may have multiple matches. Thanks to the homography refinement technique, our algorithm can select the accurate matches on the correct faces and provide a substantial improvement in the matching performance.

For this car experiment, the direct recognition rate with SIFT matches is only 18.4%, due to the large deformation and complex appearances of the objects in the dataset. The proposed method provided a more accurate rate of 65.3% with the help of the unfolding warping technique, which compensates for the occlusions with the global model map, and reduces the effects of repeatability using the context information. Currently, our algorithm is implemented in interpreted Matlab on a Pentium IV, 1.5GHz PC. A single pairwise matching, including the feature matching, homography estimation, and computation of matching cost, on average approximately takes 40 seconds.



Figure 5. Samples of car model maps.

6. Conclusions

We have proposed a novel framework for automatic object recognition. The major advantage of the proposed system is that the frontal views of the major aspects of training objects are stored and combined into a flat map, so multi-view object recognition can be posed as a simple patch mapping problem. Spatial context and topological features can thus be applied for more robust recognition. The training data in our approach consists of only 2D maps, leading to an efficient system compared to existing multi-view approaches. As the experiments show, the proposed approach can significantly improve the handling of large changes of viewpoints and performs well in solving challenging cases.

References

- R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, 2003.
- [2] V. Ferrari, T. Tuytelaars, and L. V. Gool. Integrating multiple model views for object recognition. In *CVPR*, volume II, pages 105–112, 2004.
- [3] G. E. Karras, E. Petsa, A. Dimarogona, and S. Kouroupis. Photo-textured rendering of developable surfaces in architectural photogrammetry. In *Proceedings of ISVAA*, Dublin, June 2001.
- [4] Y. Lamdan and H. Wolfson. On the error analysis of geometric hashing. In CVPR, pages 22–27, 1991.
- [5] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.
- [6] R. Nelson and A. Selinger. A cubist approach to object recognition. In *ICCV*, 1998.
- [7] F. Rothganger and *et al.* 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, pages 231–259, Mar. 2006.
- [8] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool. Towards multi-view object class detection. In *CVPR*, June 2006.
- [9] Z. Zhang and L. He. Whiteboard scanning and image enhancement. *Microsoft Tech Report*, June 2003.