VIEW-INVARIANT OBJECT RECOGNITION USING HOMOGRAPHY CONSTRAINTS

Sina Lotfian, Hassan Foroosh

University of Central Florida
Department of Computer Science
4000 Central Florida Blvd, Orlando, FL 32816

ABSTRACT

Change in viewpoint is one of the major factors for variation in object appearance across different images. Thus, view-invariant object recognition is a challenging and important image understanding task. In this paper, we propose a method that can match objects in images taken under different viewpoints. Unlike most methods in the literature, no restriction on camera orientations or internal camera parameters are imposed and no prior knowledge of 3D structure of the object is required. We prove that when two cameras take pictures of the same object from two different viewing angels, the relationship between every quadruple of points reduces to the special case of homography with two equal eigenvalues. Based on this property, we formulate the problem as an error function that indicates how likely two sets of 2D points are projections of the same set of 3D points under two different cameras. Comprehensive set of experiments were conducted to prove the robustness of the method to noise, and evaluate its performance on real-world applications, such as face and object recognition.

Index Terms— Object Recognition, View Invariance, Homography, Homology

1. INTRODUCTION

View invariance is a problem of great importance in various image analysis problems such as object recognition, retrieving arbitrary objects across different poses, human action recognition, and pose estimation, to name a few. The variation in pose can cause distortion in the feature space to the extent that many recognition algorithms may fail to recognize objects. The relationship between the rotation and translation of an object in the 3D world and the changes in the coordinates of pixels in the 2D image plane is also not trivial.

Algorithms dealing with variation in viewpoint usually make assumptions either about change in feature space caused by 3D transformations, or about the position and orientation of the cameras. Learning viewpoint manifolds [1] [2] and the latent spaces for viewpoints [3] [4] are two popular approaches taken by researchers for this problem, but they require simplifying assumptions in order to solve the problem. In this paper, a geometric approach is taken to address this problem and a solution in the most general case is provided.

We propose a template matching method based on image-domain relations in the projective space that can match objects across any pair of poses as long as the template image and the probe image have enough overlap for keypoint extraction. We prove that for one object seen by two cameras, with arbitrary intrinsic and extrinsic camera parameters, a restriction applies on the eigenvalues of the homography matrices associated with any quadruple of keypoint correspondences. By exploiting this constraint, an error function is introduced that is able to estimate how likely the provided reference and test images belong to the same object under different viewpoints.

The novelty of the paper can be summarized as follow:

- We propose a template matching method that can match the given template with any inquiry image even under a wide baseline and viewpoint changes, as long as they have overlaps.
- Unlike learning-based methods, the proposed approach does not need separate training data for each viewpoint. We also do not make any assumptions on the orientation of the cameras or their intrinsic matrices.

2. RELATED WORK

One common approach to tackle the varience in pose is to find latent spaces where the correlation between two views are maximized. Canonical correlation analysis(CCA) [3] projects the data from two views into two low dimensional subspace which are highly correlated. Sharma et al. [4] have extended CCA method so that it exploits the labels of training data to find a more discriminative projection direction. Both of the mentioned methods can exploit kernels to model non-linearity. Although methods based on latent spaces have proven to be powerful tools for both multi-view image classification and multi-modal data classification, they require learning a projection direction for every viewpoint and their ability to generalize to unseen viewpoints is limited.

Another set of solutions try to fit the given 2D images to predefined 3D shapes of objects (e.g. a face) from a single

Table 1. Desirable properties of some view-invariant recognition algorithms. If an algrothims satisfy the condition it is indicated by a check-mark.

	GUV	OSL	3DFREE	IICP
CCA [3]			✓	
GMA [4]			✓	
DPFD [11]	✓		✓	
Castillo et al [9]	✓	✓	✓	
Schels et al [8]	✓	✓		
Ours	✓	✓	✓	✓

view image [5][6]. In [7] authors propose a 3D pose normalization for face recognition in order to make it robust to variation in pose. [8] exploits 3D CAD models to detect and find the pose of objects such as bikes and cars. The use of these methods are restricted to objects with available predetermined 3D models. A rather interesting solution was proposed by [9] that does not require 3D reconstruction of the face, instead they use the cost of stereo matching as their error function. However, they make the assumption that epipolar lines are horizontal which does not hold true for the object recognition in the general case.

Ideally, we are in search of view-invariant recognition algorithms that require few training data (hopefully one shot learning) **(OSL)**, generalizable to unseen view-points **(GUV)**, work on objects without known 3D structure, **(3DFree)** and invariant to the internal camera parameters **(IICP)**. Table 1 compares the various classes of algorithms described above, in terms of these desired properties.

In this paper, we take a geometric approach to the problem of viewpoint variation. Our work is inspired by Shen et al.[10], who used homographgy constraints to recognize body pose transitions between two successive frames of two video cameras, observing human actions. Although, we are not dealing with video frames in this work, we show that the concept can be extended also to a pair of still images of a rigid object (i.e. instead of dealing with moving points in space viewed by two pairs of frames (4 images), we can extend the idea to recognizing a rigid object from two images. The key to achieve this extension is to consider quadruple of points in each camera image, instead of triplets of points in two frames of each camera The result is a rigid object recognition method that can handle unknown viewpoints and internal camera parameters.

3. PROPOSED METHOD

Given a reference image (I_r) and a query image (I_q), our goal is to determine if they belong to the same 3D object under two different viewpoints or not. First, point correspondences are extracted between I_r and I_q, and represented as $S = \{(p_{r1}, p_{q1}), (p_{r2}, q_{q2}, ...(p_{rn}, q_{rn}\}$. Such correspondences can be obtained from any keypoint extraction and matching algorithm such as SIFT[12], SURF[13] or Harris[14]. For more clarity, we use upper case letters for 3D coordinates and lower case for 2D coordinates on the image plane. We introduce an error function that in the ideal case vanishes, when there exist a unique 3D configuration of points which map to the extracted 2D keypoint correspondences. Conversely, the value of the error function increases, if such 3D configuration is not possible. Furthermore, the proposed error function is fully projective (i.e. fully defined in the image domain) and hence is invariant to camera positions and its internal parameters.

Consider the object shown in Figure 1, which consist of four 3D points $\{P_1, P_2, P_3, P_4\}$ in general positions. Two cameras (C1 and C2) that are located in two different coordinates are imaging this object as I_r and I_q. In the most general case, the two cameras would be projective with 11 degrees of freedom (i.e. different intrinsic parameters and arbitrary orientations in the 3D space). Two key observations that lead to the proposed solution are: (1) Any three of the quadruple of points define a plane in the 3D space that induces a homography between the two cameras; (2) With a quadruple of points one can obtain 4 such planes, i.e. two pairs of homographies. Each pair of homographies plays a similar role as a moving plane considered in [10], except that in our case instead of a single plane moving in time, we are considering the dual case of two planes in a rigid body. Since this case is dual to the problem considered by [10], the construct remains the same. We illustrate this using the example of Figure 1.

Let two planes π_1 (orange) in Figure 1(a) and π_2 (blue) in Figure 1(b) correspond to the triplets of 3D points $\{P_1, P_2, P_3\}$ and $\{P_1, P_2, P_4\}$, respectively. Let the corresponding image points be $p_r = \{p_{r1}, p_{r2}, p_{r3}, p_{r4}\}$ and $p_q = \{p_{q1}, p_{q2}, p_{q3}, p_{q4}\}$. We assume that no three projected points are co-linear in either views. Let also e_1 and e_2 denote the epipoles in the two images. Since epipoles are mapped across two images by the homography induced by any plane in the scene, we have

$$H_1 p_{ri} = p_{qi}, \quad i = 1, 2, 3 \tag{1}$$
$$H_1 e_1 = e_2 \tag{2}$$

$$H_2 p_{ri} = p_{qi}, \quad i = 1, 2, 4 \tag{3}$$
$$H_2 e_1 = e_2 \tag{4}$$

These equations yield a pair of homographies through which we can define $\mathcal{H} = H_1 H_2^{-1}$.

Proposition 1: H will reduce to a homology if and only if the presumed point correspondences p_r and p_q are images of the same 3D point configuration.

The immediate consequence of this observation is that two of the eigenvalues of H must be equal if the presumed point correspondences p_r and p_q are images of the same 3D

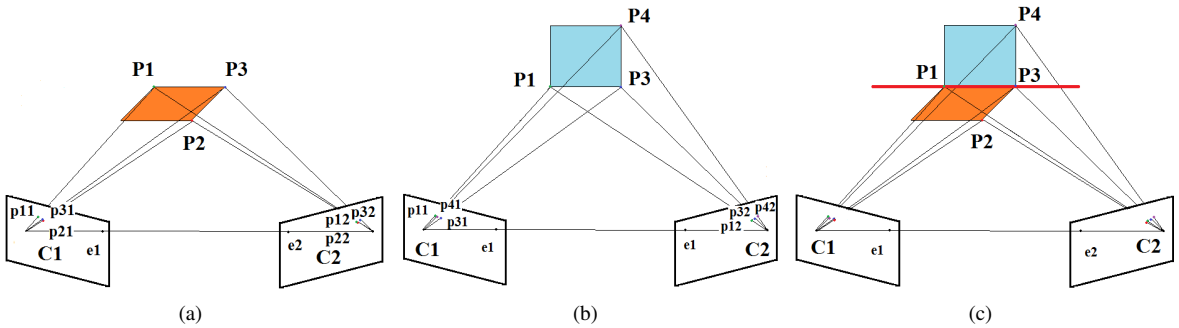


Fig. 1. Quadruple of points being viewed by two different cameras. In the part (a) you can see the plane l_1 is used to map triplet of points between two image using the homography H_1. The same rule is applied in part (b) for plane l_2 and homography H_2. Note that the red line at the intersection of two planes as well as epipoles are fixed under $H = H_1 H_2^{-1}$. Hence, it is a planar homology with two equal eigenvalues.

point configuration. This allows us to define a cost function that would make it possible to determine if a set of image points and their matching correspondences from a template image are originated from the same 3D object. Suppose we have m such template images and we establish n putative point correspondences between the query image and each reference template. One can then define $K = \binom{n}{4}$ quadruples of point correspondences, yielding a total of $2K$ matrices, $\mathcal{H}_{km}$, $k = 1, ..., 2K$, for each template $m \in \{1, ..., M\}$. Let $\epsilon_1(\mathcal{H}_{km})$ and $\epsilon_2(\mathcal{H}_{km})$ be the two closest eigenvalues of the matrix $\mathcal{H}_{km}$. Finding the optimal matching template $\hat{m}$ is then a labeling process that would be given by:

$$\hat{m} = \arg \min_{m \in \{1,...,M\}} \sum_{k=1}^{2K} \frac{|\epsilon_1(\mathcal{H}_{km}) - \epsilon_2(\mathcal{H}_{km})|}{|\epsilon_1(\mathcal{H}_{km}) + \epsilon_2(\mathcal{H}_{km})|} \quad (5)$$

4. EXPERIMENTAL RESULTS

In this section, the performance of the proposed method on both synthetic and real-world datasets is demonstrated with wide applications such as object and face recognition.

4.1. Synthetic Data

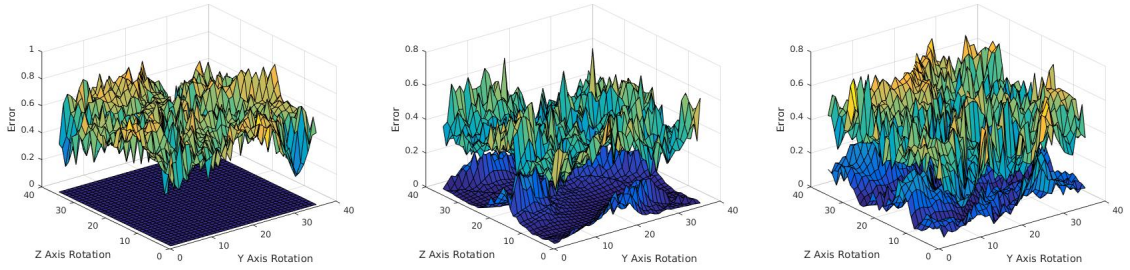
In order to understand the behavior of the error function in equation **??** in the presence of noise in key point localization, the process of projection of 3D points on the image plane is simulated using the pinhole camera model. The point clouds used for generating the synthetic objects are obtained from the BigBIRD [15] dataset, which consist of RGBD images of objects sampled on the the viewing hemisphere. Object 'Advil' is chosen as the positive example and the object 'Syrup' is chosen as negative example. It is expected that the error measure for 'Advil-Advil' pair will be lower that 'Advil-Syrup' pair.

Two cameras are used to generate synthetic images on the image plane. The first camera which is the reference camera is fixed at the world origin and is looking at the Z axis. The second camera or the test camera is moving on the viewing hemisphere. This is achieved by rotating the reference camera around Y and Z axis. Since the number of points in the cloud is over one thousand, we randomly choose 8 points as the keypoints and project only these 8 points on the image plane. The focal lengths for both cameras change randomly in the range 1000 ± 100. Then by adding Gaussian noise to the position of keypoints on the image plane, we measure the robustness of the algorithm.

In figure 2 the matching score for different viewing angles are plotted for both the matching query-template pair (the surface below) and the non-matching query-template pair (surface above). It can be observed that for the matching pair the error is almost zero, while for non-matching pairs the error is high. To find out the extent of separation between these (i.e. ability to distinguish between a correct and incorrect match), we added Gaussian noise to the position of the keypoints in the image planes. It can be observed that as the noise variance increases, the two error surfaces get closer and the ditinction between true match and a false match becomes harder. Our experiments show that we can handle noise strength of up to about $\sigma = 12$ which roughly equates the correspondences being 24 pixels off.

4.2. Real-World Data

We also tested the proposed method on real datasets, including a 3D multi-view object recognition dataset, and two multi-view face recognition datasets. The first dataset is coil-20 [16], which consists of 20 classes, each taken with the object rotated 5 degree on a turn-table. Pointing04 [17] and UMIST[18] are two multi-view face dataset used to evaluate the proposed method. UMIST consists of 575 faces from 20 different persons taken under different conditions. Pointing04



(a) $\sigma = 0$ (b) $\sigma = 6$ (c) $\sigma = 12$

Fig. 2. The error surface for different viewing angles for a) $\sigma = 0$, b) $\sigma = 6$, c) $\sigma = 12$. Without the presence of Gaussian noise the error function would vanish for same 3D point configurations between the query points and the template points. As the Gaussian noise is added to the points in the image frame, the error grows, creating a possible source of uncertainty.

Table 2. Object and face recognition dataset and their associated accuracy and size

Dataset	Accuracy	No. of Classes	Dataset Size
Coil-20	86.1	20	1440
Pointing04	77.7	15	2690
UMIST	92.5	20	565

contains 2690 face photos taken from 15 people. The Pointing04 face rotation has more degrees of freedom and images are taken with and without glasses. Keypoints are extracted using the popular SIFT [12] descriptor and matched using a nearest neighbor method. Note that unlike many methods in the literature the proposed algorithm does not assume that the coordinates of facial keypoints, such as nose and lips are given, and it only relies on the features extracted by SIFT, which may lay anywhere on the face.

Although in theory our method needs only one template per class to match two images, there has to be enough overlap between the query and reference image so that the keypoint extractor algorithm can find enough mutual keypoints in both images. Therefore, in each dataset for every class, 8 images are chosen as the templates and the rest are used as query images. For instance, in the Coil-20 dataset 8 images are taken as templates and 64 used for the test phase. The overall accuracy for all dataset are provided in 2.

5. CONCLUSION

In this paper, a new view-invariant template-matching method isn introduced that imposes no restrictions on external or internal camera parameters. The robustness of the algorithm has been tested by adding Gaussian noise to the coordinates of the keypoints on the image to simulate the behavior of error in keypoint localization. Finally, the accuracy of the method on object and face recognition was tested, producing remarkable good results.

6. REFERENCES

[1] Ahmed Elgammal and Chan-Su Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 2, pp. II–681.

[2] Amr Bakry and Ahmed Elgammal, "Untangling object-view manifold for multiview recognition and pose estimation," in *Computer Vision–ECCV 2014*, pp. 434–449. Springer, 2014.

[3] Shawe-Taylor and N. Cristianini, "Kernel methods for pattern analysis," 2004, Cambridge University Press, New York, NY, USA.

[4] Abhishek Sharma, Abhishek Kumar, Hal Daume III, and David W Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2160–2167.

[5] Dong Yi, Zhen Lei, and Stan Li, "Towards pose robust face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3539–3545.

[6] Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 605–611.

[7] Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and M Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 937–944.

[8] Johannes Schels, Joerg Liebelt, and Rainer Lienhart, "Learning an object class representation on a continuous viewsphere," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3170–3177.

[9] Carlos D Castillo and David W Jacobs, "Using stereo matching with general epipolar geometry for 2d face recognition across pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2298–2304, 2009.

[10] Yuping Shen and Hassan Foroosh, "View-invariant recognition of body pose from space-time templates," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–6.

[11] Soubhik Sanyal, Sivaram Prasad Mudunuri, and Soma Biswas, "Discriminative pose-free descriptors for face and object matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3837–3845.

[12] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[14] Chris Harris and Mike Stephens, "A combined corner and edge detector.," in *Alvey vision conference*. Citeseer, 1988, vol. 15, pp. 10–5244.

[15] Arjun Singh, James Sha, Karthik S Narayan, Tudor Achim, and Pieter Abbeel, "Bigbird: A large-scale 3d database of object instances," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 509–516.

[16] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al., "Columbia object image library (coil-20)," .

[17] Nicolas Gourier, Daniela Hall, and James L Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004, vol. 6.

[18] Daniel B Graham and Nigel M Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*, pp. 446–456. Springer, 1998.