ICSP2008 Proceedings

OBJECT RECOGNITION USING MULTI-VIEW IMAGING

Yizhou Wang, Mike Brookes and Pier Luigi Dragotti

Communications and Signal Processing Group, Electrical and Electronic Engineering Department
Imperial College London, Exhibition Road, London SW7 2AZ, England
Email: yiz.wang@imperial.ac.uk

ABSTRACT

Difficult situations such as high noise or low resolution can seriously degrade the performance of object recognition algorithms that operate on isolated images. We show that recognition performance may be improved substantially in such cases by fusing the information available from a sequence of multi-view images. In this paper we present two algorithms for object recognition based on SIFT feature points. The first operates on single images and uses chirality constraints to reduce the recognition errors that arise when only a small number of feature points are matched. The procedure is extended in the second algorithm which operates on a multi-view image sequence and, by tracking feature points in the plenoptic domain, is able to fuse feature point matches from all the available images resulting in more robust recognition.

Index Terms— Object recognition, multi-view images, local interest features, plenoptic function, SIFT.

1. INTRODUCTION

In this paper we propose an object recognition algorithm based on multi-view images and compare it with recognition from single image. The proposed algorithm exploits the structure of the multi-view data in order to propagate feature point matches from all available images onto a single reference image. We demonstrate that this approach leads to a more robust object recognition algorithm in an open-set recognition experiment.

The majority of object recognition algorithms operate on single images and are based on the matching of local interest features [1] [2]. A local interest feature extraction procedure comprises two steps: an interest point detector aimed at selecting distinctive and repeatable locations in the image, and a feature descriptor, usually a vector, that characterizes the neighborhood of the interest point. In this work we have chosen to use the scale-invariant feature transform (SIFT) described by Lowe in [3], which has consistently been shown to perform well [1].

Multi-view camera systems have attracted increasing interest in recent years [4] and many new applications that involve such systems are emerging. The data acquired by multiple cameras from any viewpoint can be parameterized by a single seven-dimensional function called the plenoptic function [5]. If we assume that cameras lie on a line and fix time and wavelength, we obtain a 3D projection of the plenoptic function known as the Epipolar-Plane Image (EPI) volume [6]. The data acquired by a multi-camera system is very structured [7]; in the case of the EPI volume, a point in the scene corresponds to a line in the plenoptic domain whose position and orientation depend on the location of the point in the scene. It is possible to exploit such structure to perform signal processing tasks in the plenoptic domain.

The paper is organized as follows: in Section 2 we introduce the feature point extraction, matching and recognition techniques that we will use, and present a single-view object recognition (SOR) algorithm. Section 3 describes our novel multi-view object recognition (MOR) algorithm. Experimental results are shown in Section 4, and we conclude in Section 5.

2. SINGLE-VIEW OBJECT RECOGNITION

Our single-view object recognition (SOR) procedure is based closely on that given by Lowe in [3], but we have modified it slightly to improve its performance. The six steps in the procedure are described briefly below; further details of the first four steps may be found in [3].

Feature Extraction: SIFT feature points are identified as the extreme values of the Difference of Gaussian (DoG) in scale and space, and are filtered to remove those with low contrast or poorly defined locations. Then an orientation is assigned to each detected point and finally, a descriptor containing 128 elements is formed for each feature point as described in [3].

Feature Matching: Feature vectors extracted from a test image are compared with those from the target dictionary using a Euclidean distance measure. A match is retained if the ratio between the distance of the best match over the second best match is sufficiently small; this is termed nearest neighbor and second nearest neighbor (NN/SNN) matching. We reject all the matches with a NN/SNN ratio greater than 0.8 as suggested in [3].

Hough Histogram Clustering: The location, scale and orientation associated with each feature point define a 2D similarity transform relating to each matched pair of features. We use a Hough histogram clustering approach to reject those matches whose similarity transform estimates are inconsistent. The estimated similarity transform parameters cast votes into a 4D histogram, consisting of translation, log scale and rotation of the similarity transform. The histogram bins are discretised to 30° for rotation, a factor of 2 for scale and 1/8 of the image size for translation. The clustering is achieved by finding the peak of this 4D histogram which will include all matches that agree on the same pose for the object. The likelihood of this pose interpretation being correct is therefore much higher.

Random Sample Consensus (RANSAC) Estimation [8] is applied to the matches filtered by NN/SNN and Hough histogram clustering to estimate a refined homography between a dictionary entry and the object in the test image. Since the Hough histogram clustering has already removed most of the outliers, RANSAC achieves good results.

Homography Validity Checking: A valid homography should map any convex region onto a non-reflected convex region of the test

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

image. When incorrect matches are obtained, it is common for the estimated homograhy to infringe this condition. This is illustrated in Figure 1(c, e, g), which shows the inverse mapping from the test image for three invalid homographies. We therefore reject any homographies that do not satisfy this chirality condition.

Interpolation and Normalized Cross-Correlation (NCC) [9]: The final step consists in interpolating the region that the homography maps from the dictionary image to the test image. Linear interpolation is applied to obtain an interpolated image with the same resolution as the dictionary entry image. We use NCC to measure the similarity between the interpolated image and the dictionary entry image, which is given by

$$r = \frac{\sum_i (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{[\sum_i (A_i - \bar{A})^2][\sum_i (B_i - \bar{B})^2]}}, \quad (1)$$

where A and B represent two images with the same size and $\bar{A}$ and $\bar{B}$ are the mean of images. For all the dictionary entries, the recognition will be made by finding the entry having the highest NCC with the interpolated image. A summary of the complete procedure for SOR is given as Algorithm 1.

Algorithm 1 Single-view Object Recognition (SOR)

1. Build a dictionary of SIFT features from the database of target objects
2. Extract SIFT features from the test image
3. Match all the features in the test image with those in the dictionary using the NN/SNN criterion
4. Use Hough histogram clustering to filter the previous NN/SNN matches
5. Apply RANSAC to find the refined homography from a dictionary entry to the object in the test image
6. Check the validity of the obtained homography
7. Interpolate the region that the homography maps from the dictionary entry image to the test image, and calculate the NCC for each entry in the dictionary
8. Choose, after a threshold checking, the entry in the dictionary with the highest NCC as the final recognition result.

Figure 1 shows the experimental results for SOR. SIFT features extracted from ten different book images are stored and indexed as the dictionary database. Figure 1(a) shows a test image which contains a copy of the dictionary entry. The matches identified in step 5 are shown as lines joining the test and dictionary images. Figure 1(b) shows the interpolated image from step 7 which resulted in a NCC of 0.69. Figure 1(c) shows the same image but this time being compared with an incorrect dictionary entry. Although some matches have been found, the interpolated image shown in Figure 1(d) results in a low NCC of 0.04 as well as an invalid homography which step 6 will reject. The SOR algorithm performs well in most of the recognition tests. However, as we can see in Figure 1(e-h), SOR can fail with noisy observations or low resolution images. In these cases SIFT does not extract a sufficient number of feature points, and NN/SNN matching together with Hough histogram clustering cannot provide enough matches. This causes the failure of SOR.

3. MULTI-VIEW OBJECT RECOGNITION

As demonstrated above, object recognition techniques based on a single-view images may fail with noisy observations or low resolution images. However, given a set of multi-view images of the same scene, more information is available and can be used to improve recognition performance. In this section we propose a novel multi-view object recognition (MOR) algorithm. Algorithm 2 gives the overview of our MOR, in which some novel techniques are exploited such as disparity-estimation and two-stage Hough histogram clustering, combined with other SOR techniques.

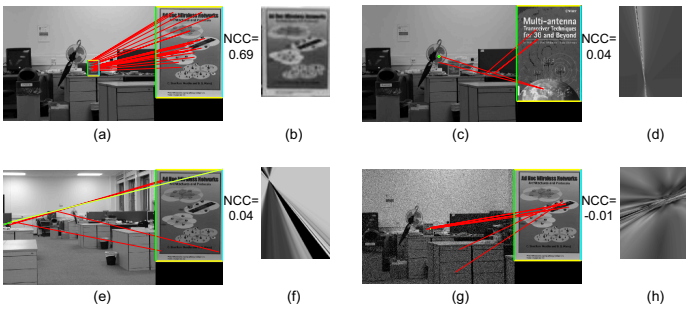

Fig. 1. Single-view object recognition. Correct dictionary entry achieves good homography and high NCC as in (a) and (b), while wrong dictionary entry obtains very low NCC in (c) and (d). Two failures are illustrated in (e) and (f) due to the low resolution test images, in (g) and (h) due to the noise, with PSNR = 40.

Algorithm 2 Multi-view Object Recognition (MOR)

1. Build the dictionary and Extract SIFT features from each multi-view test image,
2. Run Disparity-Estimation to find the disparities between the selected reference image and all other multi-view images,
   (a) Obtain the NN/SNN matches for each image.
   (b) Combine all the NN/SNN matches from all multi-view images and filter them by relaxed Hough histogram clustering.
   (c) Dynamic Programming tracking.
   (d) Find the disparities using the best tracks.
3. Perform two-stage Hough histogram clustering,
   (a) Apply 1st-stage Hough histogram clustering to filter the matches of NN/SNN for each image.
   (b) Propagate all the 1st-stage Hough matching features from all multi-view images onto the reference image using the disparities obtained from Step 2, and filter them by a 2nd-stage Hough histogram clustering.
4. Apply from Algorithm 1 steps 5 to 8 to the reference image.

3.1. Disparity-Estimation

The purpose of Disparity-Estimation is to find the disparities between a reference image and all the other multi-view images. Note that the disparities we are interested in are those of the target object, not those of other objects or background in the test images. A relaxed Hough histogram clustering method and a DP tracking technique are exploited to find the disparities as described below.

Relaxed Hough histogram clustering:

The purpose of Relaxed Hough histogram clustering is to combine the NN/SNN matches from all the multi-view images to obtain more correct matches and perform reliable and robust recognition. Recall that each match defines a similarity transform which can be used for match-filtering by Hough histogram clustering. Provided that the baseline is small compared with the distances between cameras and

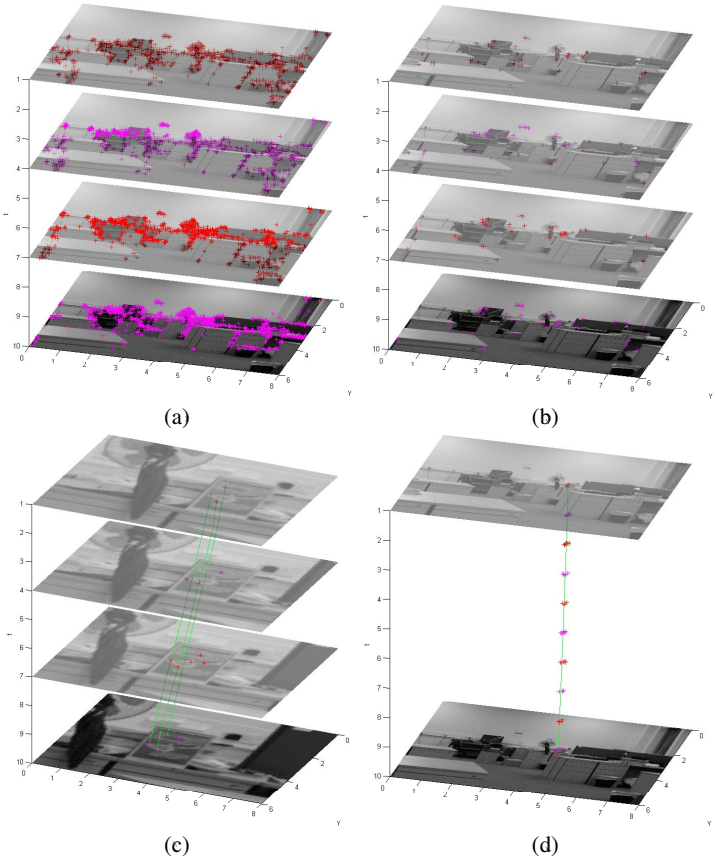

Fig. 2. 4 out of 10 natural multi-view images are illustrated as a 3D plenoptic volume to demonstrate how the disparity-estimation approach works. SIFT feature points are marked in each image with colored cross and green lines show the tracks.

the scene, it can be assumed that the similarity transforms defined by correct matches will be similar for all multi-view images. Therefore a Hough histogram clustering with relaxed bin size can be applied to filter these combined NN/SNN matches. In practice, the Hough histogram bin sizes are broadened to 45° for rotation and 1/2 of the image size for translation. Hence it is possible to cluster all the correct matches from all multi-view test images.

Disparity Estimation:

In order to track the same object feature point in multi-view images, we use a cost function comprising two components: a matching error and a geometrical error. The Euclidean distance of the feature vectors is used as the matching error. If the camera motion is known, the possible trajectories within the EPI volume [6] lie on a 1-dimensional family of curves [10]. The geometrical error component of the cost function is taken as the image-plane deviation from the best-fit curve taken from this family. Feature points can be efficiently tracked between images by using n-best Dynamic Programming (DP) [11] to minimize the cost function.

Experimental Result:

Figure 2 demonstrates the Disparity-Estimation approach in terms of multi-view images volume. The original SIFT feature points are shown in Figure 2(a) with approximately 800 features detected for each image. Figure 2(b) gives the NN/SNN matched features for the correct dictionary entry. The NN/SNN measure discards unmatched SIFT points and the number of features drops from 800 to approximately 60 for each test image. Figure 2(c) and (d) depict the matched features filtered by the relaxed Hough histogram clustering and only the correct matches on the object of interest are retained. Figure 2(c) also illustrates the DP tracking results. The final disparity-estimation result is given in (d) by averaging the 3 best tracks found in (c) by the DP tracking.

3.2. Two-stage Hough histogram clustering

Since reliable SIFT features often appear on some images and disappear on other images due to the noise and low resolution, the idea of MOR is to propagate the SIFT features of interest from all multi-view images onto a reference image. This is accomplished using the disparity obtained by DP tracking. Firstly, as in SOR, the 1st-stage Hough histogram clustering is applied to filter the NN/SNN matches in each test image. Then these filtered features from all multi-view images can be propagated onto the reference image using the disparities (i.e., the best track of the object of interest) estimated by DP. Some incorrect matches may be propagated as well, therefore a 2nd-stage Hough histogram clustering is needed to remove these incorrect propagated matches. Finally, steps 5, 6, 7 and 8 of the SOR algorithm are applied on the reference image to perform the recognition.

4. RESULTS

In this section, we present results for multi-view object recognition and compare them with the single-view method. In order to demonstrate MOR's advantage, the test data sets are chosen to be extremely hard for SOR algorithm. As in the previous SOR experiments, ten books are chosen as dictionary database and ten multi-view test images are capturing a book in a cluttered scene about 10 metres from the camera. The ten images cover a total camera displacement of about 1 metre. Figure 3 illustrates how the SOR method performs poorly with low resolution images, while the proposed MOR algorithm achieves the correct homography and a high NCC. The upper images are two examples of single-view image recognition; Image-4 was the only one resulting in a valid homography and, even so, the NCC was only 0.10. In contrast, the lower set of images are the MOR results. It can be seen that in MOR an increased number of matches are obtained by propagating the NN/SNN matches from all multi-view images using the disparities. This results in a more accurate homography and good interpolation images with a high NCC of 0.46, which gives a correct recognition result.

5. CONCLUSION AND FUTURE WORK

In the first part of this paper, we describe a single-view object recognition algorithm which extends a technique from [3]. While this algorithm works well in most cases, its performance degrades when the objects of interest are far away or when the measurements are noisy. To overcome these limitations a novel multi-view object recognition scheme is proposed. The algorithm is capable of fusing feature point matches from all the multi-view images into a single reference image. This is achieved by devising a DP method that is able to combine geometrical information such as the camera locations with the information related to each feature point in the images. Simulation results show the robustness of this algorithm to

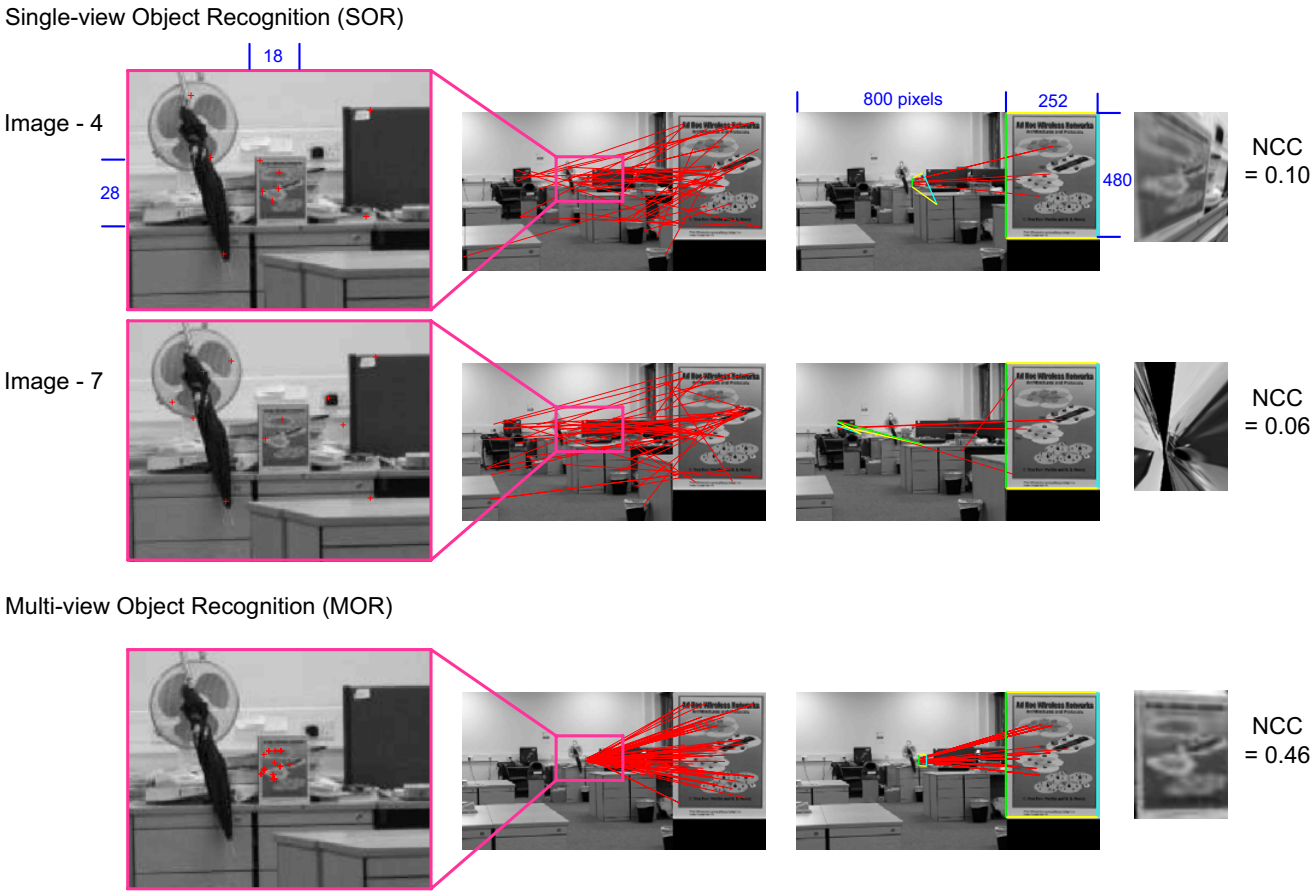

Fig. 3. Object recognition results (comparison of MOR and SOR) for low resolution images. Note that the size of interested object in the zoom-in area is 28×18 pixels, while the whole test image is 800×600 pixels, and 480×252 is the dictionary image's size. The distance between the camera and object is about 10 metres, while the total camera motion is about 1 metre.

noise and low resolution images. Images with occlusions will be taken into account in the future for comparison of MOR and SOR.

6. REFERENCES

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[2] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[3] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2003.

[4] J. Berent and P.L. Dragotti, "Plenoptic manifolds: Exploiting structure and coherence in multiview images," *IEEE Signal Processing Magazine*, vol. 24, no. 7, pp. 34–44, November 2007.

[5] E.H. Adelson and J.R. Bergen, "The plenoptic function and the elements of early vision," *Computational Models of Visual Processing*, pp. 3–20, 1991.

[6] R.C. Bolles, H.H. Baker, and D.H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[7] A. Criminisi, S.B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 51–85, 2005.

[8] M.A. Fischler and R.C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Readings in computer vision: issues, problems, principles, and paradigms*, pp. 726–740, 1987.

[9] A.L. Edwards, *Introduction to Linear Regression and Correlation*, W.H.Freeman & Co Ltd, April 1976.

[10] I. Feldmann, P. Eisert, and P. Kauff, "Extension of epipolar image analysis to circular camera movements," in *IEEE International Conference on Image Processing (ICIP)*, 2003, vol. 3, pp. 14–17.

[11] R.E. Bellman and S.E. Dreyfus, *Applied Dynamic Programming*, Princeton University Press, 1962.