Recognizing Multiple Billboard Advertisements in Videos

Naoyuki Ichimura

National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
nic@ni.aist.go.jp
http://staff.aist.go.jp/naoyuki.ichimura/

Abstract. The sponsors for events such as motor sports can install billboard advertisements at event sites in return for investments. Checking how ads appear in a broadcast is important to confirm the effectiveness of investments and recognizing ads in videos is required to make the check automatic. This paper presents a method for recognizing multiple ads. After obtaining point correspondences between a model image and a scene image using local invariants features, we separate the point correspondences of an instance of an ad by calculating a homography using RANSAC. To make the use of RANSAC feasible, we develop two techniques. First, we use the ratio of distances of descriptors to reject outliers and introduce a novel scheme to set a threshold for the ratio of distances. Second, we incorporate an evaluation on appearances of ads into RANSAC to reject the homographies corresponding to appearances of ads which are never observed in actual scenes. The detail of a recognition algorithm based on these techniques is shown. We conclude with experiments that demonstrate recognition of multiple ads in videos.

1 Introduction

The sponsors for events such as motor sports can install billboard advertisements at event sites in return for investments. Checking the positions and areas of ads in a broadcast is important to confirm the effectiveness of investments and recognizing ads in videos is required to make the check automatic. Figure 1 (a) and (b) show an example of the recognition problem. Given a model image of an ad shown in Fig. 1(a), we try to calculate the positions and areas of the instances of the ad in a scene image shown in Fig. 1(b). An issue in recognition is that ads in videos could have various appearances depending on their sizes and sites, the position, angle and zoom of a camera, and other factors. In the scene image, the changes in scales and intensities of the four instances of the ad are observed as well as occlusions. As this example demonstrates, we need to cope with deformations, illumination changes and occlusions in recognition.

Using local invariant features is a way to develop a recognition algorithm which has tolerance to deformations, illumination changes and occlusion. Local invariant features are constructed by making the following two components invariant to deformations and illumination changes: (i) local region detectors,

L.-W. Chang, W.-N. Lie, and R. Chiang (Eds.): PSIVT 2006, LNCS 4319, pp. 463–473, 2006.

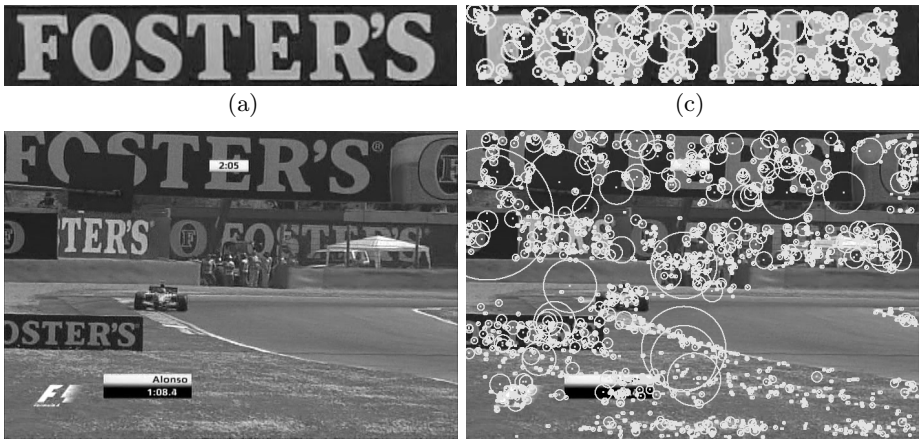


Fig. 1. An example of a recognition problem. (a) A model image of an ad. (b) A scene image obtained from a broadcast of a car race (F1). Our aim is to find the positions and areas of the instances of the ad in the scene. In the scene, the changes in scales and intensities of the four instances of the ad are observed as well as occlusions. As this example demonstrates, we need to cope with deformations, illumination changes and occlusions of ads in recognition. (c),(d) Examples of local regions. The circles represent local regions in which descriptors are computed. Only 33% of all regions are shown for the sake of clarity. Local invariant features calculated in the local regions are used to develop a recognition algorithm which has tolerance to deformations, illumination changes and occlusions of ads.

(ii) descriptors describing local regions. The circles in Fig. 1 (c) and (d) represent the examples of local regions. Using multiple local regions as shown in these figures, we can recognize ads even if they are partly occluded because the features of visible portions are available. Many different techniques for detecting and describing local regions have been developed. Local region detectors are based on interest points extraction in scale space [1,2,3,4,5,6,7,8,9,10] , region segmentation [5,11,12,13], edge detection [13,14] etc. Descriptors are derivatives of intensities [2,8], image patches obtained by region normalization [4], moment features [11,12,13], wavelet coefficients [9], histograms of gradient orientations [3,5,6,7,10,14] etc. The local invariant features constructed from these detectors and descriptors can be invariant to various transformations of a local region shape and an intensity such as similarity and affine transformations. Therefore point correspondences between a model image and a scene image are found in spite of deformations, illumination changes and occlusions.

We, however, encounter another issue after point correspondences are obtained. The issue is separation of the point correspondences of a single instance of an ad. If there are multiple instances of an ad in a scene, point correspondences of instances are mixed and the separation of them is indispensable to recognize an instance. Although many methods for object matching based on local

invariant features have been proposed [2,3,4,5,7,14], the problem of multiple instances rarely gets much attention so far.

The purpose of this paper is to develop an algorithm for recognizing multiple instances of ads. Basically, the segmentation problem to separate an instance can be treated as a model fitting problem for point correspondences with outliers (false matches) [15]. We can use a homography [16] as the general model which gives a global constraint for grouping the point correspondences of a single instance because billboards are planes in most cases. Thus to extract the point correspondences that obey a homography is equivalent to separate the point correspondences of an instance. Although a robust estimator, RANSAC (RANdom SAmple Consensus) [17], can be used to calculate a homography while rejecting outliers, it would fail if the ratio between inliers and outliers is low as noted in [7,13]. An alternative to RANSAC is Hough transformation [7], but it needs an approximation of a homography such as a similarity transformation to reduce the number of dimensions of a voting space.

In order to take advantage of a homography as the general model, we introduce two techniques for model fitting by RANSAC. First, we use the ratio of distances of descriptors to reject outliers [7,9] and introduce a novel scheme to set a threshold for the ratio of distances. The scheme ensures the reasonable number of point correspondences so that we can use RANSAC to calculate a homography. Second, we incorporate an evaluation on appearances of ads into RANSAC. Using the evaluation, we can reject the homographies corresponding to appearances of ads which are never observed in actual scenes. The evaluation is useful to reject the homographies computed from samples containing outliers that coincidentally yield a reasonable amount of votes. We will show the detail of an algorithm based on these techniques in the following sections and demonstrate by experiments that our method recognizes multiple instances of ads in various situations even if only one model image is given.

2 Recognition Algorithm

We show the proposed recognition algorithm using the following notations: The local invariant features of a scene image are represented by a set $\boldsymbol{f}_i^s = \{\boldsymbol{p}_i^s, \sigma_i^s, \boldsymbol{d}_i^s\}$, where, $\boldsymbol{p}_i^s$ indicates the position of a local region expressed in homogeneous coordinates, σ_i^s the scale in which the local region is found, $\boldsymbol{d}_i^s$ the descriptor and i index. Similarly, the local invariant features of a model image are denoted as $\boldsymbol{f}_j^m = \{\boldsymbol{p}_j^m, \sigma_j^m, \boldsymbol{d}_j^m\}$. The distance between the features is measured by the Euclidean distance between descriptors, $d_{ij} = \|\boldsymbol{d}_i^s - \boldsymbol{d}_j^m\|$.

We use Hessian-Laplace detector [7,8] to detect local regions. The detector is selected based on the results of the preliminary test of 3 local region detectors, Harris-Laplace detector [18], Hessian-Laplace detector and Difference of Gaussian (DoG) detector [3,7], using several images. The radii of circular local regions are determined as $20\sigma_i^s$ and $20\sigma_j^m$. As a descriptor, we use the gradient location-orientation histogram (GLOH) [19] based on the fact that its high invariance has been shown in the comparative experiments of [19]. For gradient

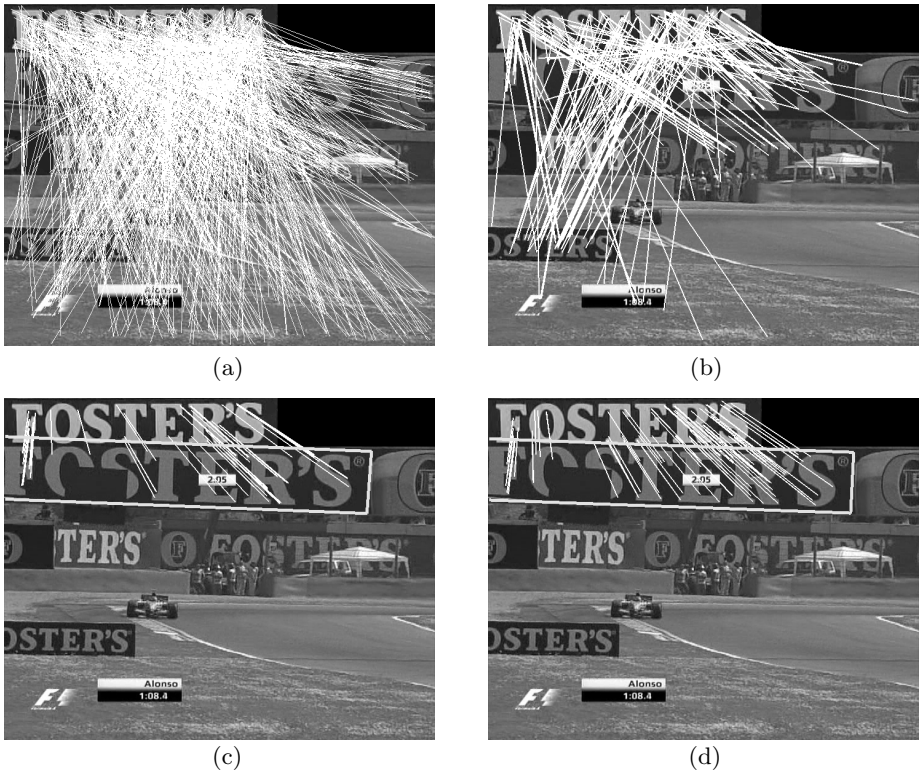


Fig. 2. An example of the recognition process. In (c) and (d), only inliers are shown as point correspondences. The results of image alignment are expressed in rectangles. (a) Matching result by the nearest neighbor method. Only 20% of all matches are shown for the sake of clarity. (b) Putative matching using Eqs.(1) and (2). Compared to (a), false matches have been reduced while keeping the inliers. However, point correspondences of the multiple instances of the ad are mixed. (c) Putative alignment by a homography. The position of the single instance is obtained by RANSAC in which an evaluation on appearances of ads is incorporated. (d) Guided matching and final alignment results. The search regions are restricted based on the results in (c), and point correspondences are obtained only from regions around the instance. Compared to Fig. 2 (c), a homography is calculated with more inliers which lead to final alignment of the instance. Regardless of the existence of the multiple instances, the point correspondences of the single instance have been separated.

locations, we use 4 bins in radial direction and 8 bins in angular direction, which results in 25 location bins. Gradient orientations are represented by 16 bins. The number of dimensions of a descriptor is $25 \times 16 = 400$.

2.1 Putative Matching with Rejection of False Matches

The nearest neighborhood method is adopted to find matches between a model image and a scene image. The feature $\boldsymbol{f}^m_{j_{1NN}}$ with the index of $j_{1NN} = \arg\min_j d_{ij}$

is matched with the feature $\boldsymbol{f}_i^s$. Figure 2 (a) shows an example of matches. As shown in this figure, there are many false matches mainly due to a background. Such false matches make the ratio between inliers and outliers low.

To reduce the number of false matches, only the point correspondences that satisfy the following equation are extracted [7,9]:

$$d_{ij_{1NN}}/d_{ij_{2NN}} < t,\ 0 \leq t \leq 1\,, \tag{1}$$

where, j_{2NN} is an index for the second nearest neighbor and t is a threshold value. The number of point correspondences extracted using Eq. (1) increases as t increases, and it reaches the maximum in the nearest neighborhood method corresponding to $t = 1$. Since the relationship between $d_{ij_{1NN}}$ and $d_{ij_{2NN}}$ depends on scene images, it is difficult to estimate the number of extracted point correspondences if we use a fixed threshold value as in [7,9]. If the number of point correspondences is too small, 4 or more inliers needed to calculate a homography may not be included. If it is too large, the ratio between inliers and outliers could be low. We actually had difficulty to set an appropriate threshold for many scenes.

To ensure the reasonable number of point correspondences, we increase t according to the following equation until the number of extracted point correspondences reaches a certain number P_{min}:

$$\begin{aligned} &t\,(k+1) = \alpha t\,(k)\ , \\ &\alpha = 1.01,\ t\,(0) = 0.80,\ k = 0, 1, 2, \ldots\ , \end{aligned} \tag{2}$$

where, k is the number of iterations, and α is the coefficient to control increase in t. By this scheme which gradually increases t, P_{min} point correspondences are ensured while trying to get the high ratio between inliers and outliers as possible as we can. Thus this scheme enable us to use RANSAC to calculate a homography. We found empirically that $P_{min} = 60$ is a good choice for many scenes. If P_{min} point correspondences cannot be extracted, our algorithm decides that there is no ad in a scene.

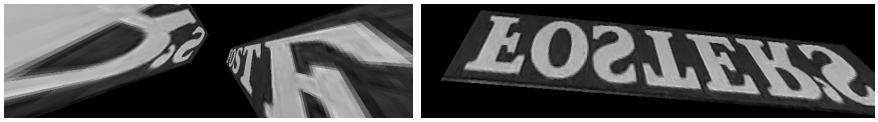
Figure 2 (b) shows an example of outlier rejection based on Eqs. (1) and (2). Compared to Fig. 2 (a), in which the nearest neighbor method is used, false matches have been reduced while keeping the inliers.

Although false matches can be reduced, the point correspondences of the instances of the ad are mixed in Fig. 2 (b). Because the point correspondences of other instances work as false matches in calculation of the homography of a certain instance, mixed point correspondences can be a cause of selecting an incorrect solution in RANSAC. The next section describes RANSAC in which an evaluation on appearances of ads is incorporated to select the correct solution.

2.2 Putative Alignment by RANSAC with Appearance Evaluation

We denote point correspondences as a set using new index k; $C = \{\boldsymbol{p}_k^m, \boldsymbol{p}_k^s\}$, $k = 1, \ldots, P$. Expressing the homography that takes each $\boldsymbol{p}_k^m$ to $\boldsymbol{p}_k^s$ as $\boldsymbol{H}$ (3×3 matrix), transformation error is defined by the following equation:

$$e_k = \|\boldsymbol{p}_k^s - \boldsymbol{H}\boldsymbol{p}_k^m\|\,,\ k = 1, \ldots, P\,. \tag{3}$$



(a) (b)

Fig. 3. Examples of appearances of ads that are rejected in RANSAC. (a) an appearance corresponding to a twisted rectangle, (b) an appearance corresponding to a reversed rectangle. Such appearances are never observed in actual scenes.

We can calculate the position and area of an instance in a scene by transforming a model image by $\boldsymbol{H}$. $\boldsymbol{H}$ can be calculated by the following RANSAC [17]:

(i) A sample comprising of 4 point correspondences is extracted randomly from the set C.

(ii) $\boldsymbol{H}$ is calculated from the sample using Direct Linear Transformation (DLT) algorithm [16] followed by non-linear optimization with the sum of transformation errors defined by Eq. (3) as an evaluation function.

(iii) Transformation errors are calculated for all point correspondences to obtain the number of inliers (votes). Inliers are the point correspondences with the errors that satisfy the following equation:

$$e_k < \varepsilon,\ k = 1, \ldots, P\,, \tag{4}$$

where, ε is the threshold value.

(iv) Processes (i) to (iii) are repeated to obtain the inliers with the maximum vote.

(v) $\boldsymbol{H}$ is calculated using the inliers obtained in (iv).

Since point correspondences of instances are mixed as in Fig. 2(b), a sample containing outliers may have the maximum vote by chance in the above algorithm. It is important to note that, in many cases, the results of transforming a model image by $\boldsymbol{H}$ computed from samples extracted from "multiple instances" yield appearances of ads which are never observed in actual scenes. Figure 3 (a) and (b) show the examples of such appearances. The major cause to take such appearances the maximum vote is that the transformation error of Eq.(3) is only criterion to select $\boldsymbol{H}$. To address the problem, the following process evaluating appearances of ads is added after (ii).

(ii') If the result of transforming a model image by $\boldsymbol{H}$ is a twisted rectangle or a reversed rectangle, the process returns to (i). If not, the process proceeds to (iii).

Twisted and reversed rectangles correspond to appearances such as Fig. 3 (a) and (b), respectively. Thus we can avoid voting by Eq. (4) for homographies corresponding to impossible appearances by the process (ii'). Twisted rectangles can be detected by whether the intersection points of lines obtained by connecting the vertexes of a model image after transformation are within the convex closure comprising of transformed vertexes. Reversed rectangle can be detected by the signed area [20] used to find the faces of polygons. Since the computations

to detect these rectangles are very efficient, the evaluation in (ii') is suited to RANSAC which requires iterations.

We calculated homographies from 10000 samples obtained from the point correspondences in Fig. 2 (b), and found 9899 homographies corresponding to twisted or reversed rectangles. Since many impossible appearances actually occur in RANSAC as seen in this case, the evaluation in (ii') is really effective for selecting the correct solution. Figure 2 (c) shows the result of putative alignment in the case of $\varepsilon = 3$ [pixel] in Eq. (4). In this figure, the lines show inliers and the rectangle is the result of transforming the model image by the homography obtained by RANSAC with appearance evaluation. The false matches and mixed point correspondences have been removed and the single instance is separated successfully.

2.3 Guided Matching

Using the result of putative alignment such as Fig. 2 (c), we can obtain point correspondences only from regions around a single instance, which excludes the effects of a background and other instances. For local invariant features of a model image, predicted positions for matching are computed using the homography $\boldsymbol{H}$ obtained in putative alignment as:

$$\hat{\boldsymbol{p}}_k^s = \boldsymbol{H}\boldsymbol{p}_k^m,\ k = 1, \ldots, P\,. \tag{5}$$

We can set circular search regions with the predicted positions as the centers and radii r_k in a scene image. The radii r_k of the search regions are determined by the scale of the features as $20\sigma_k^m$. The distances between descriptors are calculated only for local invariant features in the circular search regions and they are evaluated by Eqs. (1) and (2). If only one point correspondence is found and thus Eq. (1) cannot be evaluated, the point correspondence shall be used.

2.4 Final Alignment and Verification

Using the point correspondences obtained by guided matching, we calculate a homography again by RANSAC with appearance evaluation. Then we verify the alignment result. Circular regions are prepared for N_i inliers by the same way as guided matching. The similarity between a model image transformed by $\boldsymbol{H}$ and a scene image are measured by the normalized cross correlations of RGB channels computed within the regions. Note that the calculation of the similarity shown here can minimize the effects of occlusions, because the normalized cross correlations are calculated in the local regions instead of the entire image. If the following equation on the average value of the normalized cross correlations, $NCC_l, l = 1, \ldots, N_i$, is satisfied, the final alignment result is accepted:

$$\frac{1}{N_i}\sum_{l=1}^{N_i} NCC_l > \gamma\,, \tag{6}$$

where, γ is the threshold value. We set $\gamma = 1$.

Figure 2 (d) shows the results of final alignment. Compared to Fig. 2 (c), a homography is calculated with more inliers which lead to accurate position of the instance. The alignment result is accepted, because the average value of NCC_l in Eq. (6) is 2.3. Using the processes in the Sections 2.1 to 2.4, the single instance has been successfully separated although there are multiple instances in Fig. 2.

2.5 Termination Conditions

If the result of final alignment is accepted, the point correspondences in the area of a recognized instance (e.g., in the rectangle that shows the recognition result in Fig. 2 (d)) are removed. To recognize other instances, the processes in Sections 2.2 to 2.4 are performed for the remaining point correspondences. This procedure is repeated until one of the following termination conditions is satisfied: (a) 4 or more inliers cannot be obtained in RANSAC and (b) the condition of Eq. (6) is not satisfied. These conditions correspond to the case with no point correspondences that satisfy the global constraint and the case with an incorrect alignment result, respectively.

3 Experimental Results

We applied the proposed algorithm to videos of F1. Five ads were selected as recognition targets, and the model images shown at the top of each image in Fig. 4 were used. For each target, only one model image shown in Fig. 4 was given .

Figures 4 (a)〜(j) show the successful results. In spite of deformations, illumination changes and occlusions of the ads in these scene images, all ads were successfully recognized by separating point correspondences of each instance. These results demonstrate that our method recognizes multiple ads in various situations even if only one model image is given.

Figures 4 (k) and (l) show negative examples. Since the view point was located horizontally against the ad on the ground in front in Fig. 4 (k), the deformation of the ad is extremely large. Ads located at a far distance are observed as extremely small in size. The deformation and scaling for these ads seemed to have exceeded the range that could be compensated by the invariant property of the local features, and thus point correspondences were not obtained. In Fig. 4 (l), recognition of the top left ad failed. In this case, the degree of occlusion was too large to obtain the sufficient number of point correspondences.

As seen in the negative examples of Figs. 4 (k) and (l), recognition naturally fails if point correspondences cannot be obtained even by using local invariant features. One method for addressing such situations is to develop a local invariant feature that is better able to deal with deformations and occlusions. Another promising method is to use several model images including the deformations of ads that can be expected beforehand. We are currently working to examine these two methods.

Fig. 4. Recognition results for broadcasts of F1. The rectangles in the figures show the recognition results. Figures (a) to (j) show successful cases. In spite of deformations, illumination changes and occlusions of the ads in these scene images, all ads were successfully recognized by separating point correspondences of each instance. These results demonstrate that our method recognizes multiple ads in various situations even if only one model image is given. Figures (k) and (l) show failure cases. (k) Since the ranges of deformation and scaling that could be covered by the invariants were exceeded, point correspondences could not be obtained for the ad on the ground in front and the small ads in distance. (l) The degree of occlusion was large and recognition of the top left ad failed. Such failures may be eliminated by improving the local invariant features and using several model images including the deformations of targets that can be expected beforehand.

4 Summary

In this paper, we have presented an algorithm for recognizing multiple billboard advertisements in videos. We used the local invariant features for matching between a model image and a scene image, but false matches and mixed point correspondences appeared due to the effect of a background and multiple instances of ads. To separate the point correspondences of a single instance from the result of matching, we introduced the outlier rejection method which yields the reasonable number of point correpondences so that we can use RANSAC to calculate a homography. We also introduced RANSAC with appearance evaluation in which impossible appearances of ads are rejected. Final alignment and verification were done using the point correspondences found by guided matching based on the homography. These procedures were carried out sequentially until the termination condition was satisfied. The experimental results showed the usefulness of our algorithm. We believe that the algorithm presented here will be a good tool for several applications such as marketing research and video retrieval.

References

1. C. Harris and G. Giraudon. A combined corner and edge detector. In *Proc. 4th Alvey Vis. Conf.*, pages 147–151, 1988.
2. C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. PAMI*, 19(5):530–535, 1997.
3. D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, volume 2, pages 1150–1157, 1999.
4. M. Brown and D. Lowe. Invariant features from interest point groups. In *Proc. BMVC*, pages 253–262, 2002.
5. J. Sivic and A. Zisserman. Video google: a text retrieval approarch to object matching in videos. In *Proc. ICCV*, volume 2, pages 1470–1477, 2003.
6. M. Brown and D. Lowe. Recognising panoramas. In *Proc. ICCV*, volume 2, pages 1218–1225, 2003.
7. D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
8. K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.
9. M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *Proc. CVPR*, volume 1, pages 510–517, 2005.
10. P. Quelhas, F. Monay, J.M.Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV*, volume 1, pages 883–890, 2005.
11. F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Proc. ICCV*, volume 1, pages 636–643, 2001.
12. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, pages 384–393, 2002.
13. T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 59(1):61–85, 2004.
14. K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *Proc. BMVC*, volume 2, pages 779–788, 2003.

15. P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. R. Soc. Lond. A*, 356:1321–1340, 1998.
16. R. Hartley and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge University Press, 2nd edition, 2003.
17. M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *ACM Graphics and Image Processing*, 24(6):381–395, 1981.
18. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, volume 1, pages 525–531, 2001.
19. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. CVPR*, pages 257–264, 2003.
20. T. Moller and E. Haines. *Real-time rendering.* A.K.Peters, 2nd edition, 2002.