Shape recognition with edge-based features

K. Mikolajczyk
Dept. of Engineering Science
Oxford, OX1 3PJ
United Kingdom
km@robots.ox.ac.uk

A. Zisserman
Dept. of Engineering Science
Oxford, OX1 3PJ
United Kingdom
az@robots.ox.ac.uk

C. Schmid
INRIA Rhône-Alpes
38330 Montbonnot
France
schmid@inrialpes.fr

Abstract

In this paper we describe an approach to recognizing poorly textured objects, that may contain holes and tubular parts, in cluttered scenes under arbitrary viewing conditions. To this end we develop a number of novel components. First, we introduce a new edge-based local feature detector that is invariant to similarity transformations. The features are localized on edges and a neighbourhood is estimated in a scale invariant manner. Second, the neighbourhood descriptor computed for foreground features is not affected by background clutter, even if the feature is on an object boundary. Third, the descriptor generalizes Lowe's SIFT method [12] to edges.

An object model is learnt from a single training image. The object is then recognized in new images in a series of steps which apply progressively tighter geometric restrictions. A final contribution of this work is to allow sufficient flexibility in the geometric representation that objects in the same visual class can be recognized. Results are demonstrated for various object classes including bikes and rackets.

1 Introduction

Numerous recent approaches to object recognition [2, 12, 13, 14, 15, 20, 24] represent the object by a set of colour or grey-level textured local patches. They obtain excellent results for objects which are locally planar and have a distinctive texture [21]. However there are many common objects where texture or colour cannot be used as a cue for recognition (cf. figure 1). The distinctive features of such objects are edges and the geometric relations between them. In this paper we present a recognition approach based on local edge features invariant to scale changes. Our goal is to recognize classes of roughly planar objects of wiry components against a cluttered background. For example, bikes, chairs, ladders etc.

A very important property of our recognition approach is scale invariance [12, 14]. This enables the recognition of an object viewed from a different distance or with different camera settings. The scale invariance can locally approximate affine deformations, thereby additionally providing some immunity to out of plane rotations for planar objects. A second problem area is occlusions and background clutter. These can significantly change the appearance of features localized on object boundaries. Therefore, it is crucial to separate the foreground from the background. Since strong edges often appear on the boundaries they can be used to split the support regions before computing the descriptors.

BMVC 2003 doi:10.5244/C.17.79

1.1 Background

Our approach builds on recent object recognition methods. The idea of representing an object by a collection of local invariant patches (to avoid occlusion problems) can be traced back to Schmid and Mohr [21], where the patches were based on interest points and were invariant to rotations. Lowe [12] developed an efficient object recognition approach based on scale invariant features (SIFT). This approach was recently extended to sub-pixel/sub-scale feature localization [5]. In the context of scale invariant features Mikolajczyk and Schmid [14] developed a scale invariant interest point detector.

Recently, many authors developed affine invariant features based on the second moment matrix [2, 15, 20] or other methods [13, 24]. However, affine invariant features provide better results than scale invariant features only for significant affine deformations [15], and are not used here.

The invariance to affine geometric (and photometric transformations) reduces the already limited information content of local features. Therefore, many authors also use geometric relations between features to correctly resolve ambiguous matches. A common approach is to require that the neighbouring matches are consistent with a local estimate of a geometric transformation [18, 20, 21, 22]. This method has proved very good at rejecting false matches, and is adopted here.

Edge based method with affine [10] or projective [19] invariance, were successful in the early nineties, but fell out of favour partly because of the difficulties of correctly segmenting long edge curves. More recently recognition methods based on the statistics of local edges have been developed by Amit and Geman [1], and Carmichael and Hebert [7, 8]. The latter successfully detect objects with wiry components in cluttered backgrounds. However, many positive and negative examples are required to learn the object shape and background appearance, and there is no invariance to scale. We adopt a local edge description and incorporate the scale invariance previously only available to methods based on local appearance patches. The problem of background clutter was also handled, although manually, in the patch approach proposed by Borenstein and Ullman [4] for object segmentation.

Other related approaches using edge information are those of Belongie *et al.* [3] who use 2D shape signatures based on edges in the context of shape matching, although scale invariance and background clutter problems are not addressed in their work, and the projectively invariant shape descriptor used by Sullivan and Carlsson [25].

1.2 Overview

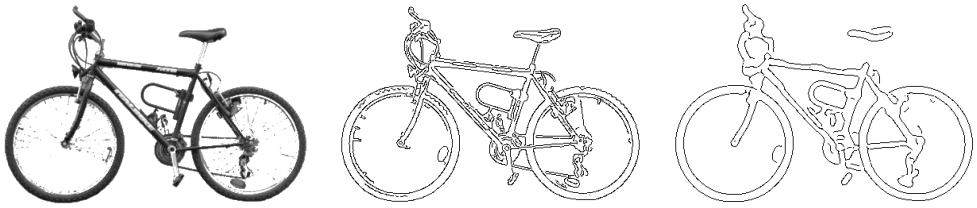
Section 2 presents the new feature detector and local edge descriptor. Section 3 describes the two stages of the recognition system: first clustering on a local transformation to reduce ambiguity, and then estimating a global (affine) transformation to detect the object in an image. In more detail, we combine an appearance distance between feature descriptors and local geometric consistency to compute the scores for point matches. The best matches with relatively few outliers are then used to vote in the Hough space of local affine transformations. The distinctive clusters in this space are used to detect and localize the objects. Section 4 gives experimental results.

2 Local features

In the following we describe our feature detector. Our objective is to determine the edge neighbourhood that is related to the scale of the object. We then show how we deal with occlusions and background clutter. Finally we present the descriptor that represents the edge shape in the point neighbourhood.

2.1 Support regions

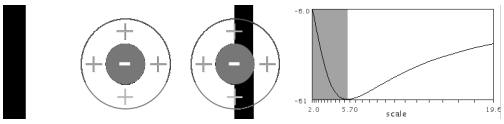
Edge features. In our task edges of low curvature and their spatial relations are very characteristic of the object. The widely used Harris [9] and DoG [12] detectors are not suitable for our purpose as the first one detects corner-like structures and the second one mostly blobs. Moreover these points are rarely localized on edges, and only accidentally on straight edges. It is well know that edge features are present at various scales and can change their appearance at different scales. Figure 1 shows the object and the edges detected with Gaussian derivatives at $\sigma = 1$ and $\sigma = 3$. The edges change their locations due to blurring, and new edges appear at different scales (cf. figure 1(b)(c)). Therefore it is crucial to build a scale-space representation to capture the possible edge appearance. To find the local features we first extract edges with a multi-scale Canny edge detector [6] using Gaussian derivatives at several pre-selected scales, with the scale interval of 1.4.



(a) (b) (c)

Figure 1: (a) Object model. (b) Edges detected at scale $\sigma = 1$. (c) Edges detected at scale $\sigma = 3$.

Scale invariance. Having computed edges at multiple scales, our goal is now to determine the size of the neighbourhood of the edge point that will be used to compute the descriptor. Several authors use the Laplacian operator for this purpose [11, 12, 14, 20]. Given a point we compute the Laplacian responses for several scales. We then select the scales for which the response attains an extremum. For a perfect step-edge the scale parameter for which the Laplacian attains an extremum is in fact equal to the distance to the step-edge. This is a well known property of the Laplacian and can be proved analytically [11]. Figure 2(a) shows an example of a ridge and 2(b) a sketch of a 2D Laplacian operator. Figure 2(c) displays the laplacian centred on one edge. Figure 2(d) shows the responses of the scale normalized Laplacian for different parameters σ. The scale trace attains a minimum for σ equal to the distance to the step-edge. There are several advantages to this approach. The first one is that we obtain characteristic scale for the edge points. This scale is related to the object scale and determines the point neighbourhood within which we capture more signal changes [14]. Figure 3 shows a few examples of point neighbourhoods selected by the Laplacian operator applied to images of different scale. Note that the feature is centred on one edge and the selected scale corresponds to the distance from the point to a neighbouring edge tangent to the circle. The edge

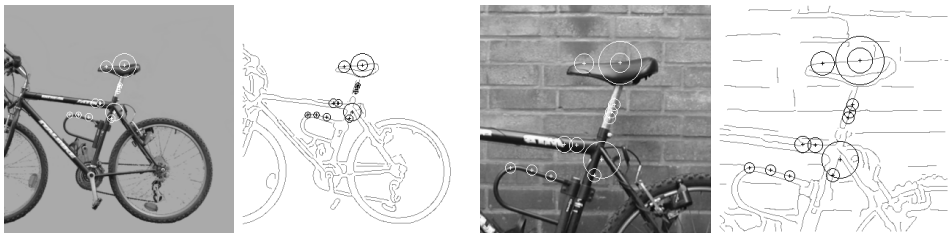


(a) (b) (c) (d)

Figure 2: Scale trace of the Laplacian localized on a 2D ridge. (a) 2D ridge. (b) Sketch of 2D Laplacian operator. (c) Laplacian localized on one edge. (d) Responses of the scaled Laplacian operator for the given location. The scale of the extremum response corresponds to the distance to the other edge.

neighbourhood is correctly detected despite the scale change and different background. A second advantage of this approach is that points which have homogeneous neighbourhood can easily be identified and rejected since they do not attain a distinctive extremum over scale. In this manner many of the edge points computed over the multiple scales are discarded.

An alternative straightforward method would be to search for tangent neighbouring edges but we found this approach less stable than the Laplacian scale selection.



(a) (b) (c) (d)

Figure 3: A few points selected by the Laplacian measure centred at the edge points. (a)(c) Images related by a scale factor of 2. (b)(d) Edges with corresponding regions. Note that the Laplacian attains an extremum when it finds another edge. The radius of the circles is equal to the selected σ.

Foreground-background segmentation. In the following we describe a new method for separating foreground and background. In the context of recognition of objects with holes and tubular components the background texture can significantly affect the descriptors such that recognition becomes impossible. To reduce the background influence, the point neighbourhood is divided into two parts separated by a chain of dominant edges, and descriptors are computed separately for each part as described below. The dominant edges are selected by locally fitting a line to the extracted edges using RANSAC. Figure 4(a) shows an example of corresponding edge points on different backgrounds. Figure 4(b) displays the gradient images and figure 4(c)(d) the selected principal edge with the neighbourhood. The tangent angle ϕ is used to obtain rotation invariance for the descriptors.

2.2 Edge Descriptors

A descriptor that captures the shape of the edges and is robust to small geometric and photometric transformations is needed for this approach. A comparative evaluation of descriptors in [16] showed that SIFT descriptors [12] perform significantly better than

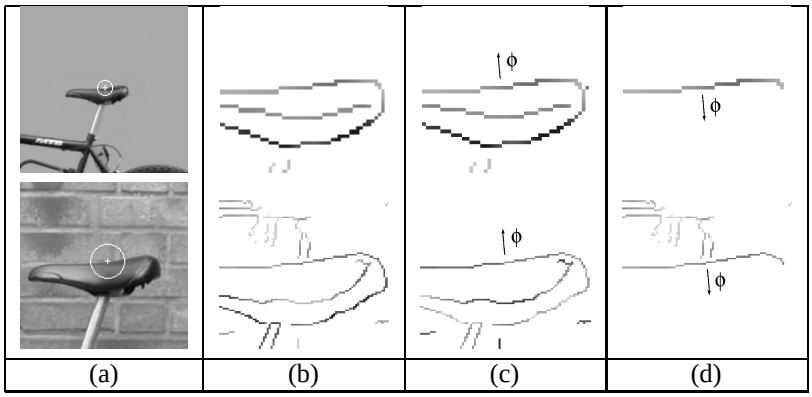


Figure 4: Background foreground segmentation. (a) Point neighbourhood, (b) Gradient edges. (c)(d) Region parts separated by the dominant edge. ϕ is the reference angle for the descriptor.

many other local descriptors recently proposed in the literature. Inspired by this result we extend the SIFT descriptor to represent the edges in the point neighbourhood. For each region part (cf. figure (a)) we build a 3D histogram of gradient values, for which the dimensions are the edge point coordinates (x, y) and the gradient orientation. The histogram bins are incremented by the gradient values at the edge points. The values are weighted by a Gaussian window centred on the region. The descriptor is built from two histograms. To compute the first we use a 2×2 location grid and 4 orientation planes (vertical, horizontal and two diagonals, cf. figure (b)). The dimension of this descriptor is 16. For the second histogram we use a 4×4 location grid and 8 orientation planes (cf. figure (c)). The dimension is 128. These two histograms are used in our coarse-to-fine matching strategy discussed in the next section. To obtain rotation invariance the gradient orientation and the coordinates are relative to the principal line separating the region (cf. figure 3(c)(d)). The descriptor of each region part contains also the points on the dominant edge. Each region part is described separately but we also use the joint descriptor to represent the whole region. To compensate for affine illumination changes we normalize each description vector by the square root of the sum of squared vector components. The similarity between the descriptors is measured with Euclidean distance.

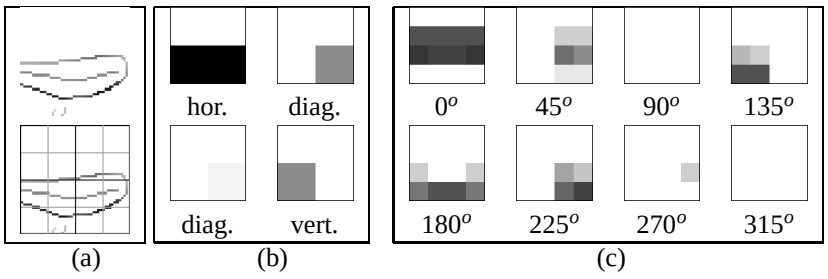


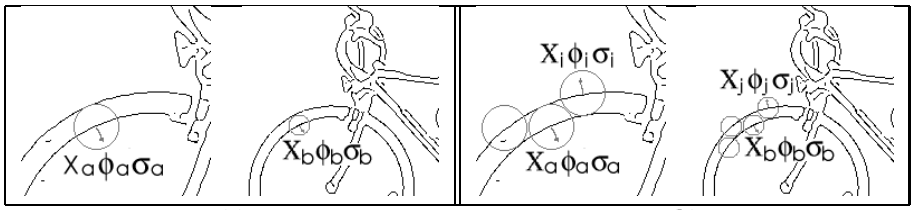
Figure 5: Edge-based local descriptor. (a) Support region and location grid. (b) Four 2×2 orientation planes. (c) Eight 4×4 orientation planes.

3 Coarse-to-fine geometric consistency

It is supposed that we have an image of the object we wish to recognize, and our task is to decide if an object of this class is present in a novel image. The recognition strategy consists of two main stages aimed at establishing matches between the model and target (novel) image. The first stage is filtering matches by taking into account the similarity of their histogram descriptors and the local geometric consistency of a similarity transformations between spatially neighbouring matches. The second stage is clustering the pose of the whole object in a coarsely partitioned affine space.

A single descriptor has small discriminatory power but a set of spatially neighbouring features which preserve the geometric relations under arbitrary viewing conditions can unambiguously identify an object. This property was successfully used in numerous approaches [5, 20, 21, 22]. Given an object to recognize, we can restrict the transformations that it can undergo. An affine transformation is a good trade-off for objects which are roughly planar when seen from a certain distance. The difficulty in estimating the transformation arises from outliers (edge mis-matches) which can often occur. An affine transformation can be locally approximated by a similarity, and this transformation requires fewer reliable inliers to compute an estimate. In our approach we can obtain an estimate of the local similarity with a single correspondence. This is because a single correspondence provides four constraints: scale change, rotation and translation. Another problem is that an exhaustive search for the transformation can be very time expensive in particular with a large number of points. To handle these problems we apply a coarse to fine strategy which rejects the outliers in successive stages.

Combined histogram distance and similarity constraint measure. In the first stage we build a 2D matrix where each entry contains a matching score for a pair of descriptors – one from the object model, the other from the target image. This score consists of two factors. The first factor is the distance between the descriptor vectors. To increase the computational speed we compute the descriptor distance for every possible pair of 16 dimensional histograms in two images. Note that a point has three descriptors corresponding to the region parts. We use only the minimum distance given by the descriptors of one point pair. We apply a distance threshold to select the matches processed in the further stages of the approach. For these matches we compute the distance d_E between the 128 dimensional histograms. The second factor contributing to the matching score



(a) (b)

Figure 6: Voting using neighbouring points. (a) Investigated matches $\mathbf{x}_a$, $\mathbf{x}_b$, where $\mathbf{x}_a$ is in the model image, and $\mathbf{x}_b$ in the target image. (b) Neighbouring features $\mathbf{x}_i$, $\mathbf{x}_j$ contribute to the score $v(\mathbf{x}_a, \mathbf{x}_b)$ (cf. equation 1).

measures the consistency of neighbouring matches to the similarity transformation estimated from the investigated pair of points. This geometric consistency is used to weight

the descriptor distance of every neighbouring point pair. The matched points $\mathbf{x}_a, \mathbf{x}_b$ give a hypothesis of the local similarity transformation between the images, where the scale change is $\sigma_{a,b} = \sigma_a/\sigma_b$ and the rotation is $\phi_{a,b} = \phi_a - \phi_b$ (cf. figure 6). The matching score for a given pair of points is:

$$v(\mathbf{x}_a, \mathbf{x}_b) = \frac{1}{1 + d_E(\mathbf{x}_a, \mathbf{x}_b)} + \sum_{i,j} \frac{\beta_{ij}\alpha_{ij}}{1 + d_E(\mathbf{x}_i, \mathbf{x}_j)} \tag{1}$$

where α and β are the penalizing functions defined by

$$\alpha_{ij} = \frac{1}{1 + 0.1\,|\phi_{a,b} - \phi_{i,j}|}\,, \qquad \beta_{ij} = \begin{cases} \sigma_{a,b}/\sigma_{i,j} & if\ \sigma_{a,b}/\sigma_{i,j} < 1 \\ \sigma_{i,j}/\sigma_{a,b} & otherwise \end{cases}$$

Points $\mathbf{x}_i$, $\mathbf{x}_j$ are spatial neighbours of points $\mathbf{x}_a$, $\mathbf{x}_b$ (cf. figure 6) within a distance $5\sigma_a$, $5\sigma_b$ respectively. The larger the difference between the scale and rotation estimated from the $\mathbf{x}_a$, $\mathbf{x}_b$ and $\mathbf{x}_i$, $\mathbf{x}_j$ pair, the smaller the weighting factors α and β. A neighbouring point contributes only once in the score for a given pair $\mathbf{x}_a$, $\mathbf{x}_b$. We use only the pairs of points which maximise the ratio $\beta_{i,j}\alpha_{i,j}/(1 + d_E(\mathbf{x}_i, \mathbf{x}_j))$. The pairs such that $\beta_{ij}\alpha_{ij} < 0.1$ are not included in the score. We use $1 + d_E$ to avoid zero in the denominator (cf. equation 1), which can happen when the distance between descriptor vectors equals zero. However, this is very rare and usually the distances between descriptors are much larger than 1.

Hough transform for affine clustering. In the second stage of our recognition approach we cluster the object pose using a voting technique in affine space. To localize the object in a target image we have to find the transformations that relates our model and the image. Note that there may be several instances of the object, and therefore several transformations may be valid. We use the generalized Hough transform to find the affine transformations between the model and the image. Given the scores computed in the previous stage we use the best matches to search for affine transformations between local groups of points. Every three non-collinear pairs of points in a limited region vote for a local affine transformation. The maximum spatial distance between points is given by the local scales of the features. We can use a larger distance since the affinity better approximates the transformations which our object undergoes.

$$A = R(\theta_1) \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} R(\theta_2) + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \tag{2}$$

We parametrize the transformation space by translations t_1, t_2 , rotation $\theta_d = \theta_1 - \theta_2$ as well as the two scalings s_1, s_2. The ratio $\sqrt{s_1 s_2}$ is the average scale of the object, and s_1/s_2 is the aspect ratio. This representation provides a possibility to discretize the scales within different ranges of values and with different intervals. This is very useful as the aspect ratio of the object changes less than its scale in real images. The votes in the transformation space are weighted by the scores obtained in the previous stage. Three pairs of matches vote for an affine transformation with a weight given by $\sum v(\mathbf{x}_a, \mathbf{x}_b)$.

Finally we search for the largest clusters in the voting space. The local maxima indicates the object location. Note that there can be more than one object in the image (cf. figure 7). We remove the overlapping clusters which appear in the transformation space at neighbouring scales and rotations.

The coarse bins used in the Hough transform allow intra-class variability in the recognition. This results in coarse accuracy of affine pose estimation, but allows small shift in relative locations of local features. Thus, a similar object with slightly different geometry also yields a distinctive cluster in the Hough space.

4 Results

To validate our approach we detect bicycles in cluttered outdoor scenes under wide viewpoint changes. Figure 7 shows several detection results. The images contain multiple instances of the object at a different scale, rotation and aspect ratio. We display the matched features consistent with the estimated affine transformation.

Figure 7(a) displays the features, up to scale 10, which are used to represent the model. There are approximately 2 500 local features. Figure 7(b) displays partially occluded object on textured background and shows all the features detected in the image. There are approximately 10 000 features. Figure 7(c) shows a hyperplane from the Hough space of affine transformations. The distinctive cluster indicates a scale 1.2, an aspect ratio 0.8, and rotation of zero. Other clusters in the space are insignificant. There are 386 inliers to the estimated affine transformation, which are displayed in figure 7(d).

Figures 7(e) and (f) present examples of multiple objects of different scales and small changes of aspect ratio. The scale changes between the model and the objects in the images are 1.9, 2.4 and 2.1, 2.2 respectively. Note that the method is not restricted to exactly the same object. The local edge descriptors convey information about the shape of the edges and not about their exact appearance in the image. The small differences in texture appearance between different instances of the object do not influence the descriptors. The use of local geometric relations between features handles the problem of ambiguous matches.

Figure 7(g) and (h) displays the objects with out of plane rotations. The average scale change is 1.3, 1.1 and the aspect ratio is 1.4, 1.1 in images (h) and (i) respectively. In our experiments with images of bicycles the aspect ratio has varied up to scale 2. This limits can be exceeded in real images but in this case the planar model is not valid any more for most of the local features detected on the bicycle. Figure 7(i) shows an example of recognition of a different object class.

Discussion and conclusions

We have presented an approach to detect a poorly textured object with holes and tubular components in highly cluttered scenes under arbitrary viewing conditions. The approach does not require learning many positive and negative examples, although this can improve the recognition performance. A single view of one positive example was used to build a generic model. The approach is not limited to the recognition of the same object. This is possible because we allow for some loss in accuracy in the descriptor representation and in the affine pose clustering. We have also demonstrated the invariance to affine transformations due to pose changes.

The edge based descriptor that we have introduced here can be used to complement texture based patch descriptors so that objects can be represented by an appropriate combination of texture and edge descriptors.

In future work we will extend the learning process to capture more controlled intra-class variations. We can learn, for example, the variation of the descriptor computed on the same parts of similar objects. We can also cluster the features into groups and perform matching between clusters rather than between single features.

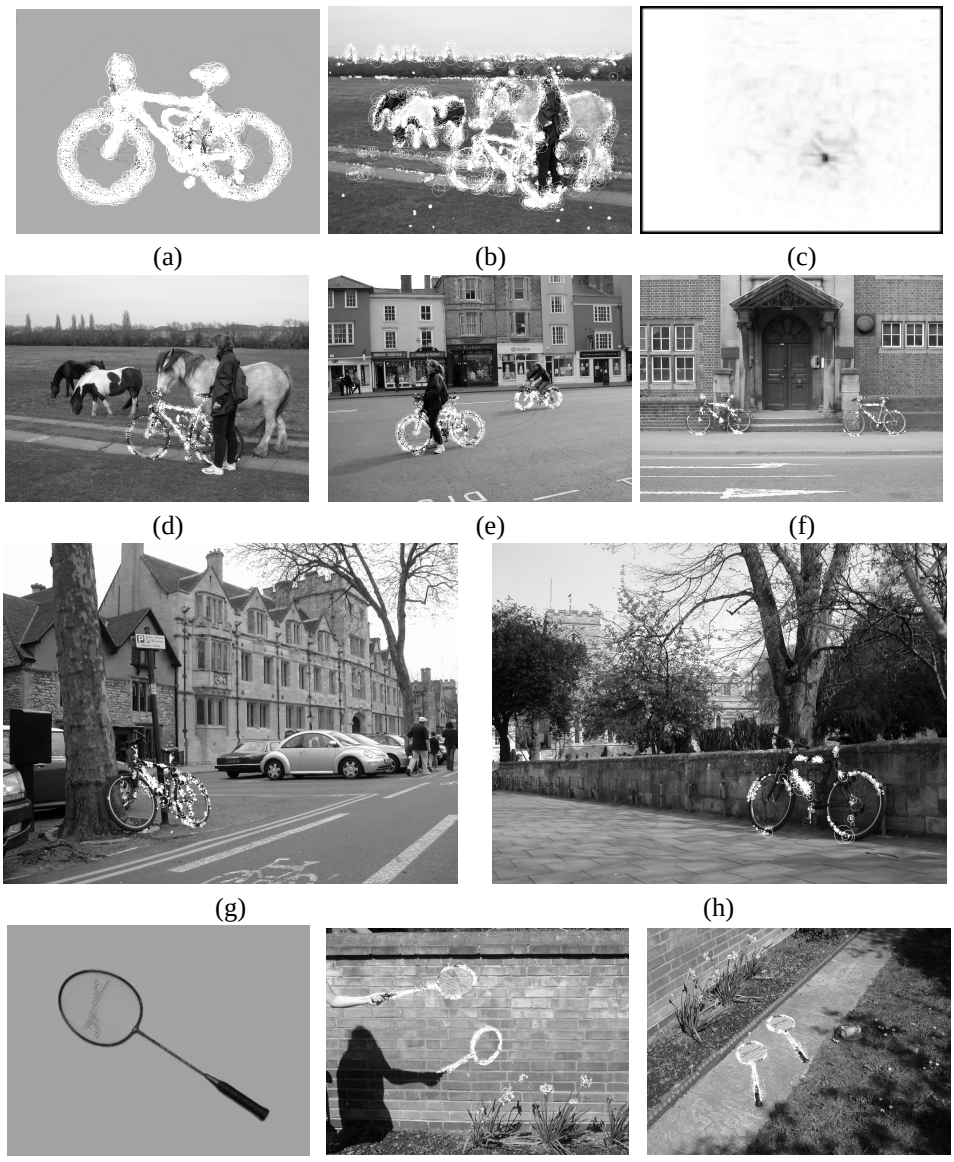


Figure 7: Examples of objects correctly recognized in images. (a) Object model, the shape is learnt from one image and the system recognizes another similar object. (b) All features detected in the target image, (c) Hyperplane of the Hough space displaying the dominant pose cluster, (d) Inliers to the estimated transformations. (e)(f) Examples with background clutter and scale changes. (g)(h) Wide viewpoint changes. (i) Another object model and matched examples.

Acknowledgements: Financial support was provided by an INRIA fellowship and EC project Vibes.

References

[1] Y. Amit and D. Geman. A computational model for visual selection. In Neural Computation, 11(7):1691–1715, 1999.

[2] A. Baumberg. Reliable feature matching across widely separated views. In CVPR, pp. 774–781, 2000.

[3] S. Belongie, J. Malik and J. Puzicha. Shape matching and object recognition using shape contexts. PAMI, 24(4):509–522, 2002.

[4] E. Borenstein and S. Ullman Class-specific, top-down segmentation. In *ECCV*, pp. 202–215, 2002.

[5] M. Brown and D. G. Lowe. Invariant features from interest point groups. In *BMVC*, pp. 253–262, 2002.

[6] J. Canny. A Computational Approach to Edge Detection. PAMI, 8(6):544–567, 1986.

[7] O. Carmichael and M. Hebert. Object recognition by a cascade of edge probes. In *BMVC*, pp. 103–112, 2002.

[8] O. Carmichael and M. Hebert. Shape-based recognition of wiry objects. In *CVPR*, pp. 401–408, 2003.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pp. 147–151, 1988.

[10] Y. Lamdan, J. T. Schwartz and H. J Wolfson. Object Recognition by Affine Invariant Matching. In *CVPR*, pp. 335–344, 1988.

[11] T. Lindeberg. *Scale-Space Theory in Computer VisionObject Recognition by Affine Invariant Matching*. Kluwer Academic Publishers, 1994.

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pp. 682-688, 1999.

[13] J. Matas, O. Chum, M. Urban and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *BMVC*, pp. 384–393, 2002.

[14] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, pp. 525–531, 2001.

[15] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, volume I, pp. 128–142, May 2002.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR*, pp. 257–264, 2003.

[17] H. Murase and S.K. Nayar Visual learning and recognition of 3D objects from appearance, In *IJCV*, 14(1):5–24, 1995

[18] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pp. 754–760, 1998.

[19] C.A. Rothwell, A. Zisserman, D.A. Forsyth and J.L. Mundy. Canonical Frames for Planar Object Recognition. In *ECCV*, pp. 757–772, 1992.

[20] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *ECCV*, pp. 414–431, 2002.

[21] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. PAMI, 19(5):530–534, May 1997.

[22] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. In *ECCV*, pp. 814–828, 2002.

[23] A. Thanyananthan, B. Stenger, P.h.S.Torr and R.Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, pp. 127–134, 2003.

[24] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, pp. 412–425, 2000.

[25] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *ECCV*, pp. 629–644, 2002.