OBJECT RECOGNITION METHODS BASED ON TRANSFORMATION COVARIANT FEATURES

Jiří Matas and Štěpán Obdržálek

Center for Machine Perception, Czech Technical University, Prague, 120 35, CZ

ABSTRACT

Methods based on distinguished regions (transformation covariant detectable patches) have achieved considerable success in a range of object recognition, retrieval and matching problems, in still images and videos. We review the state-of-the-art, describe relationship to other recognition methods, analyse their strengths and weaknesses, and present examples of successful applications.

1. INTRODUCTION

Recognition of general three-dimensional objects from 2D images and videos is a challenging task. The common formulation of the problem is essentially: given some knowledge of how certain objects may appear, plus an image of a scene possibly containing those objects, find which objects are present in the scene and where. Recognition is accomplished by matching features of an image and model of an object. The two most important issues that a method must address are the definition of a feature, and how the matching is found.

What is the goal in designing an object recognition system? Achieving *generality*, i.e. the ability to recognise any object hand-crafted adaptation to a specific task, *robustness*, the ability to recognise the objects in arbitrary conditions, and *easy learning*, i.e. avoiding special or demanding procedures to obtain the database of models. Obviously these requirements are generally impossible to achieve, as it is for example impossible to recognise objects in images taken in complete darkness. The challenge is then to develop a method with minimal constraints.

Object recognition methods can be classified according to a number of characteristics. We focus on model acquisition (learning) and invariance to image formation conditions. Historically, two main trends can be identified. In the so called geometry- or model-based object recognition, the knowledge of an object appearance is provided by the user as an explicit CAD-like model. Typically, such a model describes only the 3D shape, omitting other properties such as colour and texture. On the other end of the spectrum are the appearance-based methods, where no explicit user-provided model is required. The object representations are usually acquired through an automatic learning phase (but not necessarily), and the model typically relies on surface reflectance (albedo) properties. Recently, methods which put local image patches into correspondence emerged. Models are learned automatically, objects are represented by appearance of small local elements. Global arrangement of the representation is constrained by weak or strong geometric models.

The rest of the paper is structured as follows. In Section 2, an overview of classes of object recognition methods is given. Survey on methods which are based on matching of local features is presented in Section 3, and Section 4 describes some of their successful applications. Section 5 concludes the paper.

The authors were supported by the European Union project IST-2001-32184, by the Czech Ministry of Education project LN00B096, and by The Austrian Ministry of Education project CONEX GZ 45.535.

2. CLASSES OF OBJECT RECOGNITION METHODS

2.1 Appearance Based Methods

The central idea behind appearance-based methods is the following. Having seen all possible appearances of an object, can recognition be achieved by just efficiently remembering all of them? Could recognition be thus implemented as an efficient visual (pictorial) memory? The answer obviously depends on what is meant by "all appearances". The approach has been successfully demonstrated for scenes with unoccluded objects on black background [34]. But remembering all possible object appearances in the case of arbitrary background, occlusion and illumination, is currently computationally prohibitive.

Appearance based methods [6, 70, 20, 3, 40, 33, 68, 21, 30, 34] typically include two phases. In the first phase, a model is constructed from a set of reference images. The set includes the appearance of the object under different orientations, different illuminants and potentially multiple instances of a class of objects, for example faces. The images are highly correlated and can be efficiently compressed using e.g. Karhunen-Loeve transformation (also known as Principal Component Analysis - PCA).

In the second phase, "recall", parts of the input image (subimages of the same size as the training images) are extracted, possibly by segmentation (by texture, colour, motion) or by exhaustive enumeration of image windows over whole image. The recognition system then compares an extracted part of the input image with the reference images (e.g. by projecting the part to the Karhunen-Loeve space).

A major limitation of the appearance-based approaches is that they require isolation of the complete object of interest from the background. They are thus sensitive to occlusion and require good segmentation. A number of attempts have been made to address recognition with occluded or partial data [32, 30, 65, 5, 21, 4, 64, 20, 15, 19].

The family of appearance-based object recognition methods includes global histogram matching methods. In [66, 67], Swain and Ballard proposed to represent an object by a colour histogram. Objects are identified by matching histograms of image regions to histograms of a model image. While the technique is robust to object orientation, scaling, and occlusion, it is very sensitive to lighting conditions, and it is not suitable for recognition of objects that cannot be identified by colour alone. The approach has been later modified by Healey and Slater [14] and Funt and Finlayson [12] to exploit illumination invariants. Recently, the concept of histogram matching was generalised by Schiele [52, 51, 50], where, instead of pixel colours, responses of various filters are used to form the histograms (called then receptive field histograms).

To summarise, appearance based approaches are attractive since they do not require image features or geometric primitives to be detected and matched. But their limitations, i.e. the necessity of dense sampling of training views and the low robustness to occlusion and cluttered background, make them suitable mainly for certain applications with limited or controlled variations in the image formation conditions, e.g. for industrial inspection.

2.2 Geometry-Based Methods

In geometry- (or shape-, or model-) based methods, the information about the objects is represented explicitly. The recognition can than be interpreted as deciding whether (a part of) a given image can be a projection of the known (usually 3D) model [41] of an object.

Generally, two representations are needed: one to represent object model, and another to represent the image content. To facilitate finding a match between model and image, the two representations should be closely related. In the ideal case there will be a simple relation between primitives used to describe the model and those used to describe the image. Would the object be, for example, described by a wireframe model, the image might be best described in terms of linear intensity edges. Each edge can be then matched directly to one of the model wires. However, the model and image representations often have distinctly different "meanings". The model may describe the 3D shape of an object while the image edges correspond only to visible manifestations of that shape mixed together with "false" edges (discontinuities in surface albedo) and illumination effects (shadows).

To achieve pose and illumination invariance, it is preferable to employ model primitives that are at least somewhat invariant with respect to changes in these conditions. Considerable effort has been directed to identify primitives that are invariant with respect to viewpoint change [31, 76].

The main disadvantages of geometry-based methods are: the dependency on reliable extraction of geometric primitives (lines, circles, etc.), the ambiguity in interpretation of the detected primitives (presence of primitives that are not modelled), the restricted modelling capabilities only to a class of objects which are composed of few easily detectable elements, and the need to create the models manually.

2.3 Recognition as a Correspondence of Local Features

Neither geometry-based nor appearance-based methods discussed previously do well as defined by the requirements stated in the beginning of the paper, i.e. the *generality*, *robustness*, and *easy learning*. Geometry-based approaches require the user to specify the object models, and can usually handle only objects consisting of simple geometric primitives. They are not general, nor do they support easy learning. Appearance-based methods demanded exhaustive set of learning images, taken from densely distributed views and illuminations. Such set is only available when the object can be observed in a controlled environment, e.g. placed on a turntable. The methods are also sensitive to occlusion of the objects, and to the unknown background, thus they are not robust.

As an attempt to address the above mentioned issues, methods based on matching local features have been proposed. Objects are represented by a set of local features, which are automatically computed from the training images. The learned features are organised into a database. When recognising a query image, local features are extracted as in the training images. Similar features are then retrieved from the database and the presence of objects is assessed in the terms of the number of local correspondences. Since it is not required that all local features match, the approaches are robust to occlusion and cluttered background.

To recognise objects from different views, it is necessary to handle all variations in object appearance. The variations might be complex in general, but at the scale of the local features they can be modelled by simple, e.g. affine, transformations. Thus, by allowing simple transformations at local scale, a significant viewpoint invariance is achieved even for objects with complicated shapes. As a result, it is possible to obtain models of objects from only a few views, taken e.g. 90 degrees apart.

The main advantages of the approaches based on matching local features are summarised below.

- Learning, i.e. the construction of internal models of known objects, is done automatically from images depicting the objects. No user intervention is required except for providing the training images.
- The local representation is based on appearance. There is no need to extract geometric primitives (e.g. lines), which are generally hard to detect reliably.
- Segmentation of objects from background is not required prior recognition, and yet objects are recognised on an unknown background.
- Objects of interest are recognised even if partially occluded by other unknown objects in the scene.
- Complex variations in object appearance caused by varying viewpoint and illumination conditions are approximated by simple transformations at a local scale.
- Measurements on both database and query images are obtained and represented in an identical way.

Putting local features into correspondence is an approach that is robust to object occlusion and cluttered background in principle. When a part of an object is occluded by other objects in the scene, only features of that part are missed. As long as there are enough features detected in the unoccluded part, the object can be recognised. The problem of cluttered background is solved in a final step of the recognition process, when a hypothesised match is verified and confirmed, and false correspondences are rejected.

Several approaches based on local features have been proposed. Generally, they follow a certain common structure, which is summarised below.

Detectors. First, image elements of 'interest' are detected. The elements will serve as anchor locations in the images – descriptors of local appearance will be computed at these locations. Thus, an image element is of interest if it depicts a part of an object, which can be repeatedly detected and localised in images taken over large range of conditions. The challenge is to find such a definition of "interest", that would allow fast, reliable and precisely localised detection of such elements. The brute force alternative to the detectors is to generate local descriptors at every point. This course is obviously infeasible due to its computational complexity.

Descriptors. Once the elements of interest are found, the local image appearance in their neighbourhood has to be encoded in a way that would allow for searching of similar elements.

When designing a descriptor (also called a feature vector), several aspects have to be taken into account. First, the descriptors should be discriminative enough to distinguish between features of the objects stored in the database. Would we for example want to distinguish between two or three objects, each described by some ten odd features, the descriptions of local appearance can be as simple as e.g. four-bin colour histograms. On the other hand, handling thousands of database objects requires the ability to distinguish between a vast number of descriptors, demanding thus highly discriminative representation. This problem can be partially alleviated by using grouping, i.e. simultaneous consistent matching of several detected elements.

Another aspect in designing a descriptor is that it has to be invariant, or at least in some degree robust, to variations in an object's appearance that are not reflected by the detector. If, for example, the detector detects circular or elliptical regions without assigning an orientation to them, the descriptor must be made invariant to the orientation (rotational invariants). Or if the detector is imprecise in locating the elements of interest, e.g. having few pixel tolerance, the descriptor must be insensitive to these small misalignments. Such a descriptor might be based e.g. on colour moments (integral statistics over whole region), or on local histograms.

It follows that the major factors that affect the discriminative potential, and thus the ability to handle large object databases, of a method are the repeatability and the localisation precision of the detector.

Indexing. During learning of object models, descriptors of local appearance are stored into a database. In the recognition phase, descriptors are computed on the query image, and the database is looked up for similar descriptors (potential matches). The database should be organised (indexed) in a way that allows an efficient retrieval of similar descriptors. The character of suitable indexing structure depends generally on the properties of the descriptors (e.g. their dimensionality) and on the distance measure used to determine which are the similar ones (e.g. euclidean distance). Generally, for optimal performance of the index (fast retrieval times), such combination of descriptor and distance measure should be sought, that minimises the ratio of distances to correct and to false matches.

The choice of indexing scheme has major effect on the speed of the recognition process, especially on how the speed scales to large object databases. Commonly, though, the database searches are done simply by sequential scan, i.e. without using any indexing structure.

Matching. When recognising objects in an unknown query image, local features are computed in the same form as for the database images. None, one, or possibly more *tentative correspondences* are then established for every feature detected in the query image. Searching the database, euclidean or mahalanobis distance is typically evaluated between the query feature and the features stored in the database. The closest match, if close enough, is retrieved. These *tentative correspondences* are based purely on the similarity of the descriptors. A database object which exhibit high (non-random) number of established correspondences is considered as a candidate match.

Verification. The similarity of descriptors, on its own, is not a measure reliable enough to guarantee that an established correspondence is correct. As a final step of the recognition process, a verification of presence of the model in the query image is performed. A global transformation connecting the images is estimated in a robust way (e.g. by using RANSAC algorithm). Typically, the global transformation has the form of epipolar geometry constraint for general (but rigid) 3D objects, or of homography for planar objects. More complex transformations can be derived for non-rigid or articulated (piecewise rigid) objects.

As mentioned before, if a detector cannot recover certain parameters of the image transformations, descriptor must be made invariant to them. It is preferable, though, to have a covariant detector rather than an invariant descriptor, as that allows for more powerful global consistency verification. If, for example, the detector does not provide the orientations of the image elements, rotational invariants have to be employed in the descriptor. In such a case, it is impossible to verify that all of the matched elements agree in their orientation.

Finally, tentative correspondences which are not consistent with the estimated global transformation are rejected, and only remaining correspondences are used to estimate the final score of the match.

In the following, main contributions to the field of object recognition based on local correspondences are reviewed. The approaches follow the aforementioned structure, but differ in individual steps; in the way how are the local features obtained (detectors), and what are the features themselves (descriptors).

3. RECOGNITION AS A CORRESPONDENCE OF LOCAL FEATURES - A SURVEY

3.1 The Approach of David Lowe

David Lowe has developed an object recognition system [2, 23, 8, 7, 24, 22], with emphasis on efficiency, achieving real-time recognition times. Anchor points of interest are detected with invariance to scale, rotation and translation. Since local patches undergo more complicated transformations then similarities, a local-histogram based descriptor is proposed, which is robust to imprecisions in alignment of the patches.

Detector. The detection of regions of interest proceeds as follows:

1. Detection of scale-space peaks. Circular regions with maximal response of the difference-of-gaussians (DoG) filter, are detected at all scales and image locations. Efficient implementation exploits the scale-space pyramid. The initial image is repeatedly convolved with a Gaussian filter to produce a set of scale-space images. Adjacent scale-space images are then subtracted to produce a set of DoG images. In these images, local minima and maxima (i.e. extrema of the DoG filter response) are detected, both in spatial and scale domains. The result of the first phase is thus a set of triplets x, y and σ, image locations and a characteristic scales.
2. The location of the detected points is refined. The DoG responses are locally fitted with 3D quadratic function and the location and characteristic scale of the circular regions are determined with subpixel accuracy. The refinement is necessary, as, at higher levels of the pyramid, a displacement by a single pixel might result in a large shift in the image domain. Unstable regions are then rejected, the stability is given by the magnitude of the DoG response. Regions with the response lower than a predefined threshold are removed. Further regions are discarded which were found along linear edges, which, although having high DoG response, have unstable localisation in one direction.
3. One or more orientations are assigned to each region. Local histograms of gradient orientations are formed and peaks in the histogram determine the characteristic orientations.

The SIFT Descriptor. Local image gradients are measured at the region's characteristic scale, weighted by the distance from the region centre and combined into a set of orientation histograms. Using the histograms, small misalignments in the localisation does not affect the final description. The construction of the descriptors allows for approximately 20° 3D rotations before the similarity model fails. At the end, every detected region is represented by a 128-dimensional vector.

Indexing. To support fast retrieval of database vectors, a modification of the kD tree algorithm, called BBF (best bin first), is adopted. The algorithm is approximate in the sense that it returns the closest neighbour with high probability, or else another point that is very close in distance to the closest neighbour. The BBF algorithm modifies the kD tree algorithm to search bins in feature space in the order of their closest distance from the query location, instead of the order given by the tree hierarchy.

Verification. The Hough transform is used to identify clusters of tentative correspondences with a consistent geometric transformation. Since the actual transformation is approximated by a similarity, the Hough accumulator is 4-dimensionsional and is partitioned to rather broad bins. Only clusters with at least 3 entries in a bin, are considered further. Each such cluster is then subject to a geometric verification procedure in which an iterative least-squares fitting is used to find the best *affine* projection relating the query and database images.

3.2 The Approach of Mikolajczyk & Schmid

The approach by Schmid et al. is described in [44, 28, 56, 54, 53, 55, 27, 10]. Based on an affine generalisation of Harris corner detector, anchor points are detected and described by Gaussian derivatives of image intensities in shape-adapted elliptical neighbourhoods.

Detector. In their work, Mikolajczyk and Schmid implement affine-adapted Harris point detector. Since the three-parametric affine Gaussian scale space is too complex to be practically useful, they propose a solution which iteratively search for affine shape adaptation in neighbourhoods of points detected in uniform scale space. For initialisation, approximate locations and scales of interest points are extracted by standard multi-scale Harris detector. These points are not affine invariant because of the uniform Gaussian kernel used. Given the initial approximate solution, their algorithm iteratively modifies the shape, the scale and the spatial location of neighbourhood of each point, and converges to affine-invariant interest points. For more details see [28].

Descriptors and Matching. The descriptors are composed from Gaussian derivatives computed over the shape-normalised regions. Invariance to rotation is obtained by "steering" the derivatives in the direction of the gradient. Using derivatives up to 4th order, the descriptors are 12-dimensional. The similarity of descriptors is in first approximation measured by the Mahalanobis distance. Promising close matches are then confirmed or rejected by cross-correlation measure computed over normalised neighbourhood windows.

Verification. Once the point-to-point correspondences are obtained, a robust estimation of the geometric transformation between the two images is computed using RANSAC algorithm. The transformation used is either a homography or a fundamental matrix.

Recently, Dorko and Schmid [10] extended the approach towards object categorisation. Local image patches are detected and described by the same approach as described above. Patches from several examples of objects from a given category (e.g. cars) are collected together, and a classifier is trained to distinguish them from patches of different categories and from background patches.

3.3 The Approach of Tuytelaars, Ferrari & van Gool

Luc van Gool and his collaborators developed an approach based on matching of local image features [73, 75, 11, 72, 71, 74, 69]. They start with detection of elliptical or parallelogram image regions. The regions are described by a vector of photometricaly invariant generalised colour moments, and matching is typically verified by the epipolar geometry constraint.

Detector. Two methods for extraction of affinely invariant regions are proposed, yielding geometry- and intensity-based regions. The regions are affine covariant, they adapt their shape to the underlying intensity profile, in order to keep on representing the same physical part of an object. Apart from the geometric invariance, photometric invariance allows for independent scaling and offsets for each of the three colour channels. The region extraction always starts by detecting stable anchor points. The anchor points are either Harris points [13], or local extrema of image intensity. Although the detection of Harris points is not really affine invariant, as the support set over which is the response computed is circular, the points are still fairly stable under viewpoint changes, and could be precisely localised (even to subpixel accuracy). Intensity extrema, on the other hand, are invariant to any continuous geometric transformation and to any monotonic transformation of the intensity, but are not localised as accurately. On colour images, the detection is performed three times, separately on each of the colour bands.

Descriptors and Matching. In the case of geometry-based regions, each of the regions is described by a vector of 18 generalised colour moments [29], invariant to photometric transformations. For the intensity-based regions, 9 rotation-invariant generalised colour moments are used. The similarity between the descriptors is given by the Mahalanobis distance, correspondences between two images are formed from regions with the distance mutually smallest. Once corresponding regions have been found, the cross-correlation between them is computed as a final check before accepting the match. In the case of the intensity-based regions, where the rotation is unknown, the crosscorrelation is maximised over all rotations. Good matches are further fine-tuned by non-linear optimisation: the crosscorrelation is maximised over small deviations of the transformation parameters.

Verification. The set of tentative correspondences is pruned by both geometric and photometric constraints. The geometric constraint basically rejects correspondences contradicting the epipolar geometry. Photometric constraint assumes that there is always a group of corresponding regions that undergo the same transformation of intensities. Correspondences that have singular photometric transformation are rejected. Recently, a growing flexible homography approach was presented, which allows for accurate model alignment even for nonrigid objects. The size of the aligned area is then used as a measure of the match quality.

3.4 The LAF Approach of Matas et al.

The approach of Matas et al. [25, 37, 26, 36] starts with detection of Maximally Stable Extremal Regions. Affine covariant local coordinate systems (called Local Affine Frames, LAFs) are then established, and measurements taken relative to them describe the regions.

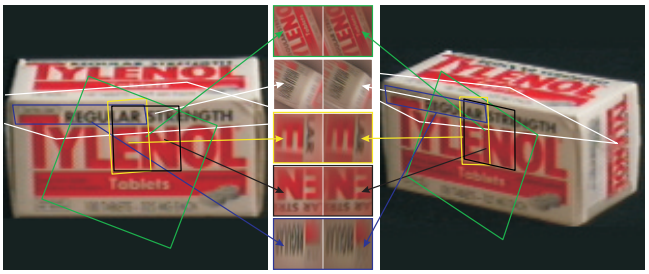

Figure 1: Examples of correspondences established between frames of a database image (left) and a query image (right).

Detector. The *Maximally Stable Extremal Regions* (MSERs) were introduced in [25]. The attractive properties of MSERs are: 1. invariance to affine transformations of image coordinates, 2. invariance to monotonic transformation of intensity, 3. computational complexity almost linear in the number of pixels and consequently near real-time run time, and 4. since no smoothing is involved, both very fine and coarse image structures are detected. Starting with contours

of the detected region, local frames (coordinate systems) are constructed in several affine covariant ways. Affine covariant properties of covariance matrix, bi-tangent lines, and line parallelism are exploited. As demonstrated in Figure 1, local affine frames facilitate normalisation of image patches into a canonical frame and enable direct comparison of photometricaly normalised intensity values, eliminating the need for invariants.

Descriptor. Three different descriptors were used. The first is directly the intensities of the local patches [37, 26, 36]. The intensities are discretised into $15 \times 15 \times 3$ rasters, yielding 675-dimensional descriptors. The size is discriminative enough to distinguish between a large amount of database objects, yet coarse enough to be tolerant to decent misalignments in the frame localisation. Second type of descriptor employs the discrete cosine transformation, which is applied to the discretised patches [38]. The number of low frequency DCT coefficients that are kept in the database is used to adapt the preference of descriptor discriminativity against the localisation tolerance. Finally, rotational invariants were used [25].

Verification. In the wide-baseline stereo problems, the correspondences are verified by robustly selecting only these conforming to the epipolar geometry constraint. For object recognition it is typically sufficient to approximate the global geometry transformation by a homography with flexible toleration increasing towards the object boundaries.

3.5 The Approach of Zisserman et al.

A. Zisserman and his collaborators developed strategies for matching of local features mainly in the context of the wide-baseline stereo problem [43, 42, 48, 45, 46]. Recently they presented an interesting work relating image retrieval problem and text retrieval [63, 47, 49]. They introduced an image retrieval system, called VideoGoogle, which is capable of processing and indexing full-length movies.

Detectors and Descriptors. Two types of detectors of local image elements are employed. One is the shape-adapted elliptical regions by Mikolajczyk and Schmid, as described in Section 3.2, second the Maximally Stable Extremal Regions from Section 3.4. Representation of the local appearance is realised by the SIFT descriptors introduced by David Lowe (see Section 3.1). Knowing that a motion video sequence is being processed, noisy and unstable regions can be eliminated. The regions detected in each frame of the video are tracked using a simple constant velocity dynamic model and correlation. Any region which does not survive for more than three frames is rejected. The estimate of the descriptor for a region is then computed by averaging the descriptors throughout the track.

Indexing and Matching. The descriptors are grouped into clusters, based on their similarity. In analogy to stop-lists in text retrieval, where common words, like 'the', are ignored, large clusters are eliminated. When a new image is observed, each descriptor of the new image is matched only against representants of individual clusters. Selection of the nearest cluster immediately generates matches for all frames of the cluster, throughout the whole movie. The exhaustive comparison with every descriptor of every frame is thus avoided. The similarity measure, used for both the clustering and the closest cluster determination, is given by the Mahalanobis distance of the descriptors.

Verification. Video frames are first retrieved using the frequency of matched descriptors, and then re-ranked based on a measure of spatial consistency of the correspondences. The matched regions provide affine transformation between the query and the retrieved image, so a point to point correspondence is locally available. A search area of each match is defined by few nearest neighbours. Other regions which also match within this area casts a vote for that frame. Matches with no support are rejected. The final rank of the frame is determined by the total number of votes.

3.6 Other Related Work

Scale Saliency by Kadir & Brady

Kadir and Brandy presented an algorithm [17] that define image regions salient if they are unpredictable in some specific feature-space, i.e. if exhibiting high entropy with respect to a chosen representation of local appearance. The approach offers a more general model of feature saliency compared with conventional techniques, which define saliency only with respect to a particular set of properties, chosen in advance.

In its basic form, the algorithm is invariant only to similarity transformations (thence the name 'scale' saliency; only the scale of circular regions is estimated on top of their locations). Recently, an affine extension to the scale selection was presented [18], capable of detecting elliptical regions. The modified saliency measure is then a function of three parameters representing the affine deformation, instead of the single one for the scale.

Local PCA, approaches of Jugessur and Ohba

As discussed in Section 2.1, global PCA (principal component analysis) based methods are sensitive to variations in the background behind objects of interest, changes in the orientation of the objects, and to occlusion. Ohba and Ikeuchi [39] and Jugessur and Dudek [16] propose an appearance-based object recognition method robust to variations in the background and occlusion of a substantial fraction of the image.

In order to apply the eigenspace analysis to recognition of partially occluded objects, they propose to divide the object appearance into small windows, referred to as "eigen windows" [39], and to apply eigenspace analysis to them. Like in other approaches exploiting local appearance, even if some of the windows are occluded, the remaining are still effective and can recover the object identity and pose.

In addition to robustness to occlusions, Jugessur and Dudek [16] also address the problem of rotation invariance. The proposed solution is to compute the PCA not on the intensity patches, but rather in frequency domain of the windows represented in polar coordinates.

The Approach of Selinger & Nelson

The object recognition system developed by Nelson and Selinger at the University of Rochester exploits a four-level hierarchy of grouping processes [35, 59, 61, 58, 57, 60]. The system architecture is similar to other local feature-based approaches though a different terminology is used. Inspired by the Gestalt laws and perceptual grouping principles, a four-level grouping hierarchy is built, where higher levels contains groups of elements from lower levels.

The hierarchy is constructed as follows. At the fourth highest level, a 3D object is represented as a topologically structured set of flexible 2D views. The geometric relation between the views is stored here. This level is used for geometric reasoning, but not for recognition. Recognition takes place at the third level, the level of the component views. In these views the visual appearance of an object, derived from a training image, is represented as a loosely structured combination of a number of local context regions. Local context regions (local features) are represented at the second level. The regions can be thought of as local image patches that surround first level features. At the first level are features (detected image elements) that are the result of grouping processes run on the image, typically representing connected contour fragments, or locally homogeneous regions. Only

Figure 2: Examples of corresponding query (left columns) and database (right columns) images from the ZuBuD dataset. The image pairs exhibit occlusion, varying illumination and viewpoint and orientation changes.

"strong" first level features are used as keys, around which context patches (the second level) are constructed.

Efficient recognition is achieved by using a database implemented as an associative memory of keyed context patches. An unknown keyed context patch recalls associated hypotheses for all known views of objects that could have produced such context patch. These hypotheses are processed by a second associative memory, indexed by the view parameters, which partitions the hypotheses into clusters that are mutually consistent within a loose geometric framework (these clusters are the third level groups). The looseness is obtained by tolerating a specified deviation in position, size, and orientation. The bounds are set to be consistent with a given distance between training views (e.g. approximately 20 degrees). The output of the recognition stage is a set of third level groupings that represent hypotheses of the identity and pose of objects in the scene, ranked by the total evidence for each hypothesis.

4. APPLICATIONS

Approaches matching local features have been experimentally shown to obtain state-of-the-art results. Here we present few examples of the adressed problems. Results are demonstrated using the approach of Matas et al. [37, 36, 38], although comparable results have been shown by others.

Figure 3: Image retrieval on FOCUS dataset: query localisation results. query images, database images, and query localisations

Object Recognition. In object recognition experiments, Columbia Object Image Library (COIL-100) [1], or more often its subset COIL-20, has been widely used, and for comparison purposes has become a de facto standard benchmark

training views/object	18	8	4	2	1
total test views	5400	6400	6800	7000	7100
LAFs	99.9%	99.4%	94.7%	88%	76%
SNoW/edges [77]	94.1%	89.2%	88.3%	-	-
SNoW/intensity [77]	92.3%	85.1%	81.5%	-	-
Linear SVM [77]	91.3%	84.8%	78.5%	-	-
Spin-Glass MRF [9]	96.8%	88.2%	69.4%	58%	50%
Nearest Neighb. [77]	87.5%	79.5%	74.6%	-	-

Table 1: COIL-100: Recognition rate (rank 1), in comparison to appearance based methods

Figure 4: An example of matches established on a wide-baseline stereo pair.

dataset. COIL-100 is a set of colour images of 100 different objects, where 72 images of each object were taken at pose intervals of 5°. The objects are unoccluded and on uncluttered black background. Such a configuration is benign for appearance-based methods. Table 1 compares recognition rates achieved by the LAF approach with the rates of several appearance-based object recognition methods. Results are presented for five experimental set-ups, differing in the number of training views per object. Decreasing the number of training views increases demands on the method's generalisation ability, and on the insensivity to image deformations. The LAF approach performs best in all experiments, regardless of the number of training views. For only four training views, the recognition rate is almost 95%, demonstrating the remarkable robustness to local affine distortions.

Image retrieval. The retrieval performance of the LAF method was evaluated on the FOCUS dataset, containing 360 colour high-resolution images of advertisements scanned from magazines. The task was to retrieve adverts for a given product, given a query image of the product logo. Examples of query logos, retrieved images, and visualised localisations of the logos are depicted in Figure 3.

Another challenging retrieval problem involved recognition of buildings in urban scenes. Given an image of an unknown building, taken from an unknown viewpoint, the algorithm was to identify the building. The experiments were conducted on a set of images of 201 different buildings. The dataset was provided by ETH Zurich and is publicly available [62]. The database contains five photographs of every of the 201 buildings, and a separate set of 115 query images is provided. Examples of corresponding query and

database images are shown in Figure 2. The LAF method achieved 100% recognition rate in rank 1.

Video retrieval. The problem of retrieval of video frames from full-length movies was addressed in [63]. Local descriptors were computed on key frames and stored into database. To reduce the otherwise enormous database size, descriptors were clustered according to their similarity. Impresive real-time retrieval was achieved for a closed system, i.e. for the case of query images originating from the movie itself.

Wide baseline stereo matching. For a significant variety of scenes the epipolar geometry can be computed automatically from two (or possibly more) uncalibrated images, showing the scene from significantly different viewpoints. The role of the matching in the wide-baseline stereo problem is to provide corresponding points, i.e. the points which in the two images represent identical element of the 3D scene. Correspondences found in a difficult stereo pair are shown in Figure 4.

5. CONCLUSIONS

In this paper we analysed and reviewed object recognition methods, focusing on these based on matching of local features. We presented a literature survey, and stated the relationship to other recognition methods. Examples of successful applications in realistic conditions were presented, demonstrating the strengths of the local methods. The applications included recognition of household objects in a database of 100 objects, recognition of buildings in a database of 200 buildings, retrieval of advertisements and the wide-baseline stereo matching.

The challenging and interesting problem of object categorisation was not covered.

REFERENCES

[1] Columbia object image library. http://www.cs.columbia.edu/CAVE.

[2] J.S. Beis and D.G. Lowe. Indexing without invariants in 3d object recognition. *PAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1000–1015, October 1999.

[3] Peter N. Belhumeur, Joao Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *ECCV (1)*, pages 45–58, 1996.

[4] H. Bischof and A. Leonardis. Robust recognition of scaled eigenimages through a hierarchical approach. In *CVPR98*, pages 664–670, 1998.

[5] H. Bischof, H. Wildenauer, and A. Leonardis. Illumination insensitive eigenspaces. In *ICCV01*, pages I: 233–238, 2001.

[6] T.E. Boult, R.S. Blum, S.K. Nayar, P.K. Allen, and J.R. Kender. Advanced visual sensor systems (1998). In *DARPA98*, pages 939–952, 1998.

[7] M. Brown and D. Lowe. Invariant features from interest point groups. In *BMVC02*, 2002.

[8] M. Brown and D.G. Lowe. Recognising panoramas. In *ICCV03*, pages 1218–1225, 2003.

[9] B. Caputo, J. Hornegger, D. Paulus, and H. Niemann. A spin-glass markov random field for 3-d object recognition. Technical Report LME-TR-2002-01, Lehrstuhl für Mustererkennung, Institut für Informatik, Universität Erlangen-Nürnberg, 2002.

[10] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV03*, pages 634–640, 2003.

[11] V. Ferrari, T. Tuytelaars, and L.J. Van Gool. Wide-baseline multiple-view correspondences. In *CVPR03*, pages I: 718–725, 2003.

[12] B.V. Funt and G.D. Finlayson. Color constant color indexing. *PAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, May 1995.

[13] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey88*, pages 147–152, 1988.

[14] G. Healey and D.A. Slater. Using illumination invariant color histogram descriptors for recognition. In *CVPR94*, pages 355–360, 1994.

[15] M. Jogan and A. Leonardis. Robust localization using eigenspace of spinning-images. In *OMNIVIS00*, 2000.

[16] D. Jugessur and G. Dudek. Local appearance for robust object recognition. In *Computer Vision and Pattern Recognition (CVPR'00)*, pages 834–840, June 2000.

[17] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, November 2001.

[18] T. Kadir and M. Brady. Scale saliency : A novel approach to salient feature and scale selection. In *International Conference Visual Information Engineering 2003*, pages 25–28, 2003.

[19] J. Krumm. Eigenfeatures for planar pose measurement of partially occluded objects. In *CVPR96*, pages 55–60, 1996.

[20] A. Leonardis and H. Bischoff. Dealing with occlusions in the eigenspace approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 453–458, June 1996.

[21] Ales Leonardis and Horst Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding: CVIU*, 78(1):99–118, 2000.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 2004.

[23] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

[24] D.G. Lowe. Local feature view clustering for 3d object recognition. In *CVPR01*, pages I:682–688, 2001.

[25] Jiří Matas, Ondřej Chum, Martin Urban, and Tomáš Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 384–393, 2002.

[26] Jiří Matas, Štěpán Obdržálek, and Ondřej Chum. Local affine frames for wide-baseline stereo. In *ICPR02*, August 2002.

[27] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV01*, pages I: 525–531, 2001.

[28] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV02*, page I: 128 ff., 2002.

[29] F. Mindru, T. Moons, and L. Van Gool. Recognizing color patterns irrespective of viewpoint and illumination. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 368–373, 1999.

[30] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *International Conference on Computer Vision (ICCV'95)*, pages 786–793, Cambridge, USA, June 1995.

[31] J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. Book, 1992.

[32] H. Murase and S.K. Nayar. Image spotting of 3d objects using parametric eigenspace representation. In *SCIA95*, pages 325–332, 1995.

[33] P. Navarrete and J. Ruiz del Solar. Comparative study between different eigenspace-based approaches for face recognition. *Lecture Notes in Computer Science*, 2275:178–??, 2001.

[34] S.K. Nayar, S.A. Nene, and H. Murase. Real-time 100 object recognition system. In *ARPA96*, pages 1223–1228, 1996.

[35] R.C. Nelson and A. Selinger. Perceptual grouping hierarchy for 3d object recognition and representation. In *DARPA98*, pages 157–163, 1998.

[36] Štěpán Obdržálek and Jiří Matas. Local affine frames for image retrieval. In *The Challenge of Image and Video Retrieval (CIVR2002)*, July 2002.

[37] Štěpán Obdržálek and Jiří Matas. Object recognition using local affine frames on distinguished regions. In *The British Machine Vision Conference (BMVC02)*, September 2002.

[38] Štěpán Obdržálek and Jiří Matas. Image retrieval using local compact dct-based representation. In *DAGM 2003: Proceedings of the 25th DAGM Symposium*, pages 490–497, 9 2003.

[39] K. Ohba and K. Ikeuchi. Detectability, uniqueness and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048, September 1997.

[40] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, WA, June 1994.

[41] A.R. Pope. Model-based object recognition: A survey of recent research. In *Univ. of British Columbia*, 1994.

[42] Philip Pritchett and Andrew Zisserman. Matching and reconstruction from widely separated views. *Lecture Notes in Computer Science*, 1506:78–85, 1998.

[43] Philip Pritchett and Andrew Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760, 1998.

[44] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR03*, pages II: 272–277, 2003.

[45] F. Schaffalitzky and A. Zisserman. Geometric grouping of repeated elements within images. In *BMVC98*, 1998.

[46] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *ICCV01*, pages II: 636–643, 2001.

[47] F. Schaffalitzky and A. Zisserman. Automated scene matching in movies. In *CIVR02*, pages 186–197, 2002.

[48] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or 'how do i organize my holiday snaps?'. In *ECCV02*, page I: 414 ff., 2002.

[49] F. Schaffalitzky and A. Zisserman. Automated location matching in movies. *CVIU*, 92(2-3):236–264, November 2003.

[50] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV96*, pages I:610–619, 1996.

[51] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR96*, 1996.

[52] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal on Computer Vision*, 36(1):31–50, January 2000.

[53] C. Schmid. Constructing models for content-based image retrieval. In *CVPR01*, pages II:39–45, 2001.

[54] C. Schmid and R. Mohr. Combining grey value invariants with local constraints for object recognition. In *CVPR96*, pages 872–877, 1996.

[55] C. Schmid and R. Mohr. Image retrieval using local characterization. In *ICIP96*, page 18A1, 1996.

[56] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, May 1997.

[57] A. Selinger and R.C. Nelson. A perceptual grouping hierarchy for appearance-based 3d object recognition. *CVIU*, 76(1):83–92, October 1999.

[58] A. Selinger and R.C. Nelson. Improving appearance-based object recognition in cluttered backgrounds. In *ICPR00*, pages Vol I: 46–50, 2000.

[59] A. Selinger and R.C. Nelson. Appearance-based object recognition using multiple views. In *CVPR01*, pages I:905–911, 2001.

[60] A. Selinger and R.C. Nelson. Minimally supervised acquisition of 3d recognition models from cluttered images. In *CVPR01*, pages I:213–220, 2001.

[61] Andrea Selinger. *Analysis and Applications of Feature-Based Object Recognition*. PhD thesis, Dept. of Computer Science, University of Rochester, New York, 2001.

[62] Hao Shao, Tomáš Svoboda, and Luc Van Gool. ZuBuD — Zurich Buildings Database for Image Based Recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology, March 2003.

[63] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV03*, pages 1470–1477, 2003.

[64] D. Skocaj, H. Bischof, and A. Leonardis. A robust pca algorithm for building representations from panoramic images. In *ECCV02*, page IV: 761 ff., 2002.

[65] D. Skocaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *ICCV03*, pages 1494–1501, 2003.

[66] M.J. Swain and D.H. Ballard. Indexing via color histograms. In *Ph. D.*, 1990.

[67] M.J. Swain and D.H. Ballard. Color indexing. *International Journal on Computer Vision*, 7(1):11–32, November 1991.

[68] Daniel L. Swets and Juyang Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[69] A. Turina, T. Tuytelaars, T. Moons, and L.J. Van Gool. Grouping via the matching of repeated patterns. In *ICAPR01*, pages 250–259, 2001.

[70] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[71] T. Tuytelaars, L. Van Gool, L. D'haene, and R. Koch. Matching of affinely invariant regions for visual servoing. In *International Conference on Robotics and Automation*, pages 1601–1606, 1999.

[72] T. Tuytelaars, A. Turina, and L.J. Van Gool. Noncombinatorial detection of regular repetitions under perspective skew. *PAMI*, 25(4):418–432, April 2003.

[73] T. Tuytelaars and L.J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC00*, 2000.

[74] Tinne Tuytelaars. *Local Invariant Features for Registration and Recognition*. PhD thesis, University of Leuven, ESAT - PSI, 2000.

[75] Tinne Tuytelaars and Luc J. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Visual Information and Information Systems*, pages 493–500, 1999.

[76] I. Weiss. Geometric invariants and object recognition. *International Journal on Computer Vision*, 10(3):207–231, June 1993.

[77] M. H. Yang, D. Roth, and N. Ahuja. Learning to Recognize 3D Objects with SNoW. In *ECCV 2000*, pages 439–454, 2000.