Object Recognition using Local Affine Frames on Maximally Stable Extremal Regions

Štěpán Obdržálek and Jiří Matas

Center for Machine Perception, Czech Technical University Prague

1 Introduction

Viewpoint-independent recognition of objects is a fundamental problem in computer vision. Recently, considerable success in addressing the problem has been achieved by approaches based on matching of regions detected by processes that are quasi-invariant to viewpoint changes [16, 20, 19, 30, 28]. Such methods represent objects by sets of regions described by invariants computed from local measurements. The representation is learned from training images without manual intervention. During recognition, the same representation is built for the test image. The recognition problem is then formulated as a search for a geometrically consistent set of correspondences of regions in the test image and in one of the training (database) images. The search proceeds in two steps. First, a tentative set of correspondence is selected on the basis of similarity of local invariants. In a seconds step, a subset of the tentative correspondences that satisfies a certain geometric constraint, e.g. epipolar geometry, is sought. The confidence in the presence of an object is expressed as a function of the matched correspondences. Since it is not required that all local features match, the approaches are robust to occlusion and cluttered background. Since the framework is based on region-to-region correspondences, recognition also achieves localisation.

This chapter describes a method which represents objects by sets of measurements defined in local coordinate systems (*local affine frames*, LAFs) that are established on affine-covariant regions [21]. The LAFs are constructed by exploiting multiple affine-covariant procedures that take the detected regions as an input. Assuming local planarity and adequacy of the affine approximation of the geometric changes induced by a movement of a perspective camera, any photometrically normalized image measurement expressed in local affine frame coordinates is viewpoint-invariant. Appearance of the objects is thus represented by local patches with shapes and locations given by the object-centred affine coordinate systems. The need for further processing of local image measurements to obtain invariant description, such as rotational or differential invariants, is eliminated. The structure of the proposed object recognition method is summarised in Algorithm 1 (the first four steps are visualised in Figure 1).

Affine coordinate systems cannot be constructed directly from interest points (e.g. [10, 14, 19]), or elliptical regions [29, 20], since neither resolves all six degrees of freedom which an affine transformation possesses. A detector of more complex image regions is required. Such regions are e.g. obtained by various segmentation techniques [9, 1] or the maximally stable extremal region (MSER) detector [18], which we exploit. MSER regions are of general, data-dependent shape, i.e. complex enough to provide sufficient constraints to define affine frames. They are connected, arbitrarily shaped, possibly nested, and do not cover the entire image, i.e. they do not form a partitioning. The formal definition of MSERs and a detailed description of the extraction algorithm is given in [18]. MSER performance evaluation and comparison to other detectors can be found in [21].

The rest of the chapter is structured as follows. In Section 2, an overview and a taxonomy of affine-covariant constructions of local coordinate systems (frames) are presented, the affine covariance of the constructions is proven, and computation issues discussed. Section 3 describes the process of geometric and photometric normalisation of local appearance. A method for forming local region-to-region correspondences is described in Section 4. In Section 5, state of the art results are reported on publicly available object recognition tests (COIL-100, ZuBuD, FOCUS). Changes of scale and illumination conditions, out-of-plane rotation, occlusion, local anisotropic scaling, and 3D translation of the viewpoint are all present in the test problems.

Algorithm 1: Structure of the proposed MSER-LAF method

1. For every database and query image, compute affine-covariant regions of data-dependent shape.
2. Construct local affine frames (LAFs) on the regions using several affine-covariant constructions.
3. Generate intensity representations of local image patches normalised according to the local affine frames. Photometrically normalise the patches.
4. Establish tentative correspondences between frames of query and database images. Compute similarity between the patches, select most similar pairs.
5. Find a globally consistent subset of the correspondences. Infer the presence and location of the objects.

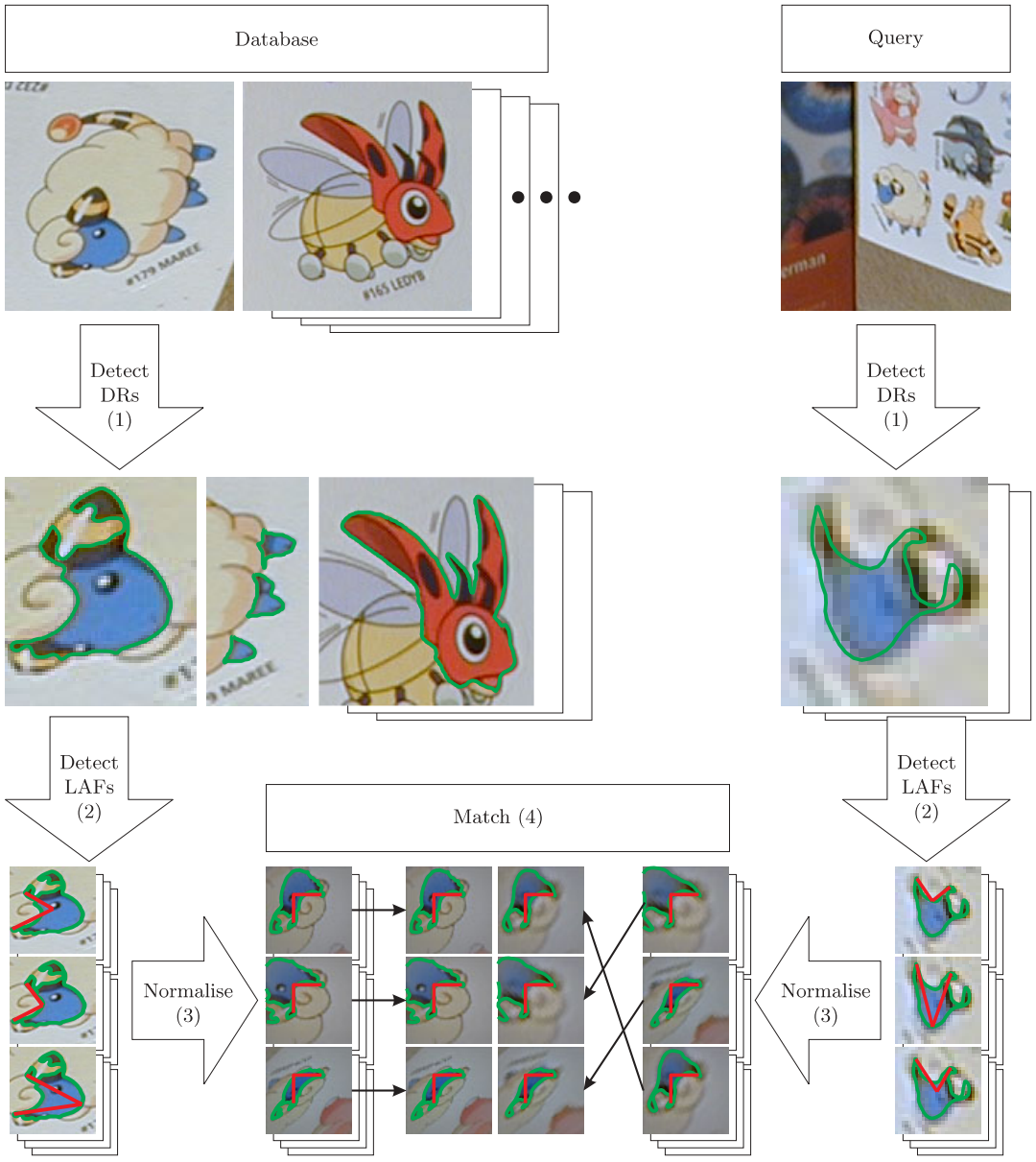


Fig. 1. Structure of the proposed MSER-LAF object recognition method

2 Local Affine Frames

2.1 Geometric Primitives Covariant with Affine Transformations

A two-dimensional affine transformation possesses six degrees of freedom. Thus, to determine an affine transformation, six independent constraints, e.g. given by a correspondence of three non-collinear points, have to be found. The constraints are derived from various affine-covariant geometric primitives detected on image regions of sufficiently complex shape. In particular, we use directions (providing a single constraint), 2D positions (providing two constraints), and the covariance matrix of a 2D shape (providing three constraints).

Figure 2 presents an overview of the affine-covariant primitives. From regions output by a detector (left top corner), other regions are affine-invariantly derived (rectangular boxes). Individual primitives (elliptical boxes) are then computed, the flow of the computation is indicated by arrows. We divide the primitives into three categories:

planar region Ω	is a contiguous subset of $\mathbb{R}^2$.
affine transformation	is a map $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form $F(\mathbf{x}) = \mathsf{A}^\mathsf{T}\mathbf{x} + \mathbf{t}$, for all $\mathbf{x} \in \mathbb{R}^n$, where A is a linear transformation of $\mathbb{R}^n$, assumed non-singular.
centre of gravity μ	of a region Ω is $\mu = \frac{1}{\lvert\Omega\rvert}\int_\Omega \mathbf{x} d\Omega$, where $\lvert\Omega\rvert$ is the area of the region.
covariance matrix	(matrix of second-order central algebraic moments) of a region Ω is a 2×2 matrix defined as $\Sigma = \frac{1}{\lvert\Omega\rvert}\int_\Omega (\mathbf{x}-\mu)(\mathbf{x}-\mu)^\mathsf{T} d\Omega$.
convex hull	of a geometric object (such as a point set or a polygonal region) is the smallest convex set S containing that object. A set S is convex if whenever two points P and Q are inside S, then the whole line segment PQ is also in S, or, equivalently, a set S is convex if it is exactly equal to the intersection of all the half planes containing it.
bitangent	is a line that is tangent to a curve at two distinct points. Bitangents contain segments of the convex hull that bridge concavities.
curvature κ	of a planar curve is defined by $\kappa = \frac{d\Phi}{ds}$ where Φ is the tangential (or turning) angle, and s is segment length. The curve is called convex in areas of positive curvature and concave in areas of negative curvature.
inflection point	is a point on a curve at which the sign of the curvature κ (i.e. the convexity of the curve) changes.

Table 1. Definition of terms

Constructions derived from region shape only. The *centre of gravity* μ (Figure 2 i) of a region (the vector of first order algebraic moments) provides two constraints, i.e. resolves translation. The symmetric 2×2 *covariance matrix* Σ (ii), the matrix of second central algebraic moments, gives 3 constraints. Together, the centre of gravity and the covariance matrix fix the affine transformation up to an unknown rotation. Normalisation by the covariance matrix (see Figure 4) therefore allows for affine-invariant measurement of distances, angles and curvatures. From these we derive the points of *extremal distance* to the centre of gravity (iii) (2 constraints) and the points of *curvature extrema* (iv) (2 constraints).

Another group of shape-derived primitives is obtained on *concavities* (v) (4 constraints for the two tangent points). Given a bitangent, the point on the region boundary *farthest from the bitangent line* (vi) is defined affine-covariantly (2 constraints). A significant property of bitangents is their locality, i.e. they do not depend on correct detection of the whole region. If, for example, two regions get connected due to discretisation in one of the images, constructions based on integral characteristics, as is the centre of gravity or the covariance matrix, are incorrect, while concavities may be unaffected.

Finally, we exploit points of *curvature inflections* (vii), i.e. points where the shape changes from concave to convex or vice-versa (2 constraints), *straight line segments* (viii) of the region boundary, and *third order algebraic moments* [12] (ix).

Constructions derived from image intensities. Several constraints can be derived from pixel values inside a region or in its neighbourhood. After normalisation by the covariance matrix, directions based on

orientations of gradients (x), obtained for example as peaks of gradient histogram [16], or the direction of *dominant texture periodicity* (xi), determine the unknown rotation. *Extrema of R, G, B components* (xii), or of any scalar function of RGB values provide 2 constraints.

Constructions derived from topology of regions. Finally, mutual configurations of regions are considered, i.e. *nested regions*, *neighbouring regions*, *holes* and *incident regions*. Region concavities and holes can be considered as distinguished regions of their own, and the computation of all of the constructions can be recursively applied. On the other hand, neither holes nor concavities have to be considered as part of the region, i.e. a convex hull can be substituted for the region, without loosing the affine invariance.

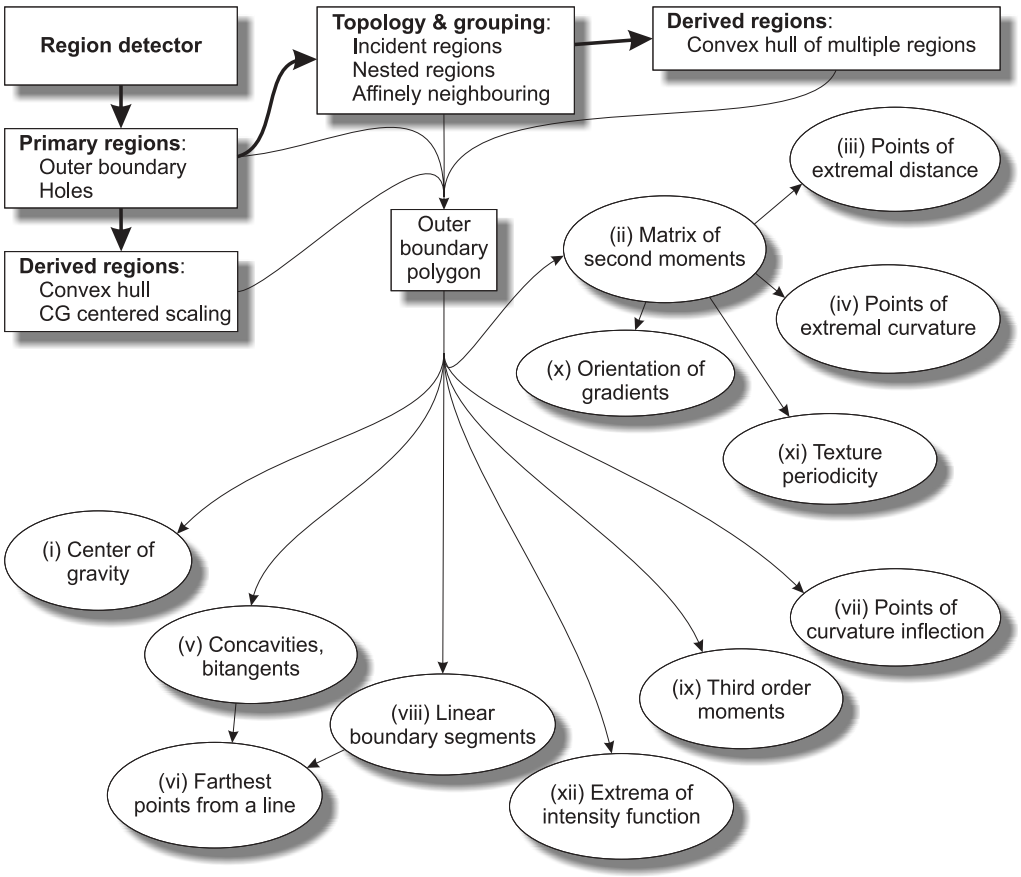
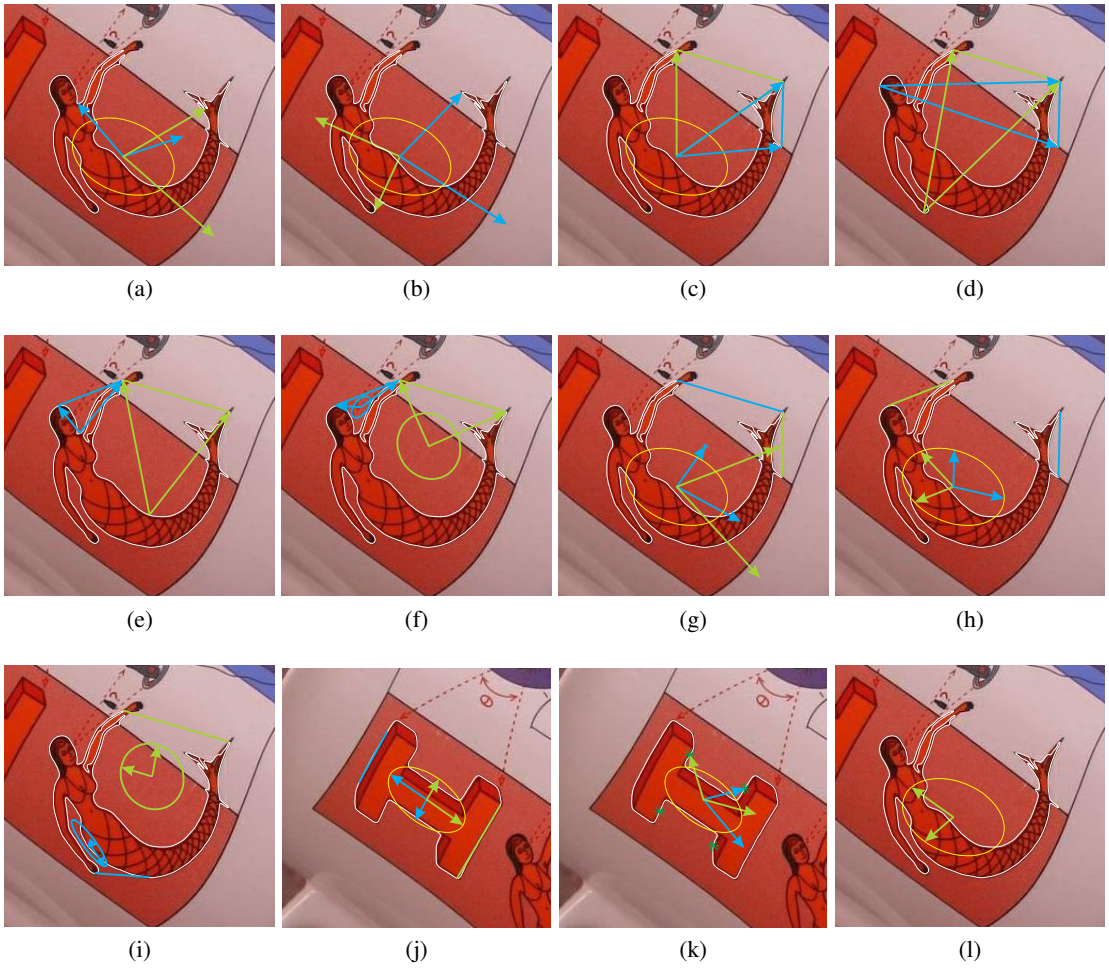


Fig. 2. Overview of affine-covariant primitives. Rectangular blocks represent regions, detected or derived, and elliptical blocks represent the primitives. The numbering refers to Sections 2.1, 2.2, and to Figure 3. Local affine frames are constructed by combining the primitives.

2.2 Details on Detection of the Geometric Primitives

A region is a connected sets of image pixels. A polygonal representation is constructed from its outer boundary. To reduce effects of discretisation, the polygons are smoothed by applying a Gaussian kernel to individual coordinates [22]. The regions are henceforth treated as simple (non-intersecting) polygons with non-integral coordinates, region holes are treated separately.

Computation of region characteristics. Let us have a polygon Ω with n vertices. Let us denote x_i and y_i the ith vertex. The polygon is closed, so $x_0 = x_n$, $y_0 = y_n$. The algorithms for computation of region area (zero order algebraic moment), centre of gravity (first order algebraic moments) and covariance matrix (second order central algebraic moments) follow the standard algorithm for computation of the area



Geometric primitive	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)
Centre of gravity of region (i)	×	×	×				×	×		×	×	×
Covariance matrix of region (ii)	×	×					×	×		×	×	×
Curvature minima* (iv)	×											
Curvature maxima* (iv)		×										
Tangent points of concavity (v)			×	×	×	×						
Farthest point on the contour (vi)				×								
Farthest point on the concavity (vi)					×							
Centre of gravity of concavity (i)						×	×		×			
Covariance matrix of concavity (ii)									×			
Direction of bitangent (v)								×	×			
Direction CoG to inflection point (vii)											×	
Direction of linear segment (viii)										×		
Direction from third-order moments (ix)												×

Fig. 3. Examples of local affine frames of different types. The table indicates which affine-covariant primitives were combined to obtain the frames.

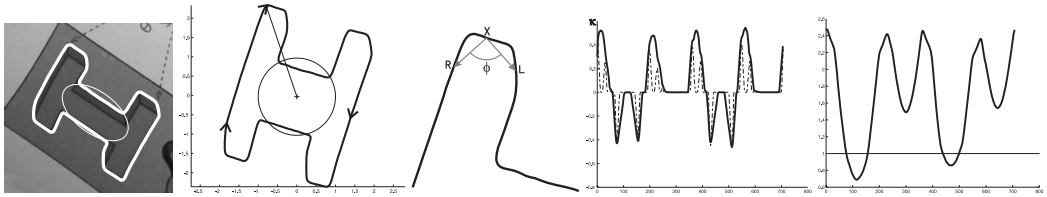
* Affine-covariant localisation of curvature extrema requires prior shape normalisation by covariance matrix.

of a non-intersecting polygon, where the area is incrementally updated for vertical strips bounded by x coordinates of two neighbouring vertices:

$$\mu_{pq} = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \int_{0}^{y_{i-1}+(y_i-y_{i-1})\frac{x-x_{i-1}}{x_i-x_{i-1}}} x^p y^q \,\mathrm{d}y\,\mathrm{d}x, \qquad \text{resp.} \tag{1}$$

$$\mu'_{pq} = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \int_{0}^{y_{i-1}+(y_i-y_{i-1})\frac{x-x_{i-1}}{x_i-x_{i-1}}} (x-\mu_{10})^p (y-\mu_{01})^q \,\mathrm{d}y\,\mathrm{d}x. \tag{2}$$

The region area is $|\Omega| = \mu_{00}$, the centre of gravity (i) is $\mu = (\mu_{10}, \mu_{01})^\mathsf{T}$, and the covariance matrix (ii) is $\Sigma = \begin{pmatrix} \mu'_{20} & \mu'_{11} \\ \mu'_{11} & \mu'_{02} \end{pmatrix}$.



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

Fig. 4. Shape normalisation by the covariance matrix. (a) detected region, (b) the region shape-normalised to have an identity covariance matrix, (c) curvature estimation, (d) curvature of the normalised shape, (e) distances to the centre of gravity.

Once the covariance matrix is computed, the region shape is normalised so that the covariance matrix of the resulting shape equals to the identity matrix. This is achieved by transforming every polygon vertex by the inverse of Cholesky decomposition of the covariance matrix, i.e. by $\mathsf{A} = (\mathrm{chol}(\Sigma))^{-1}$. The effect is illustrated in Figure 4, a detected region (a) is transformed into its normalised shape (b).

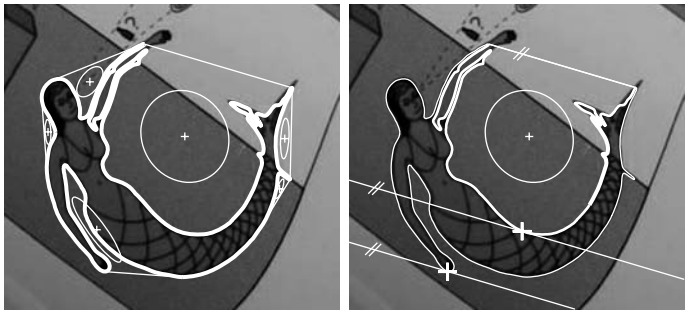
Shape normalisation, together with the position of the centre of gravity of the region, fixes the affine transformation up to an unknown rotation. The rotation is determined from local extrema of curvature (iv) or from contour points of extremal distance to the centre of gravity (iii). The computation of the curvature proceeds as follows: For each vertex X, two segments $l = \overline{\mathsf{X}L}$ and $r = \overline{\mathsf{X}R}$ of length a are spanned in opposite directions along the polygon boundary (see Figure 4 (c). The Cosine of the angle ϕ is $\cos\phi = \frac{l_x r_x + l_y r_y}{|l||r|}$, and the curvature κ is estimated as

$$\text{curvature } \kappa = s\frac{1+\cos\phi}{2}, \qquad \text{where } s = \left\{ \begin{array}{r|l} 1 & \text{if } l_x r_y - l_y r_x > 0 \\ -1 & \text{otherwise} \end{array} \right. \tag{3}$$

which ranges from -1 to 1, equals to 0 for straight segments, and is negative for concave and positive for convex curvatures. An example of the curvature values is shown in Figure 4 (d). The segment length a controls the scale at which is the curvature computed. Since the regions are shape and scale normalised, a is of a fixed value and need not be adapted to individual regions. Figure 4 (d) shows curvatures computed for two different values of a, $a = 0.5$ (thick line) and $a = 0.2$ (thin dashed line). In the experiments we use $a = 0.5$. Figure 4 (e) shows distances of vertices on the normalised contour to the centre of gravity of the region.

Inflection points (vii) are detected by an approach similar to that of computation of the local curvature. Two segments of the length a are spanned from every polygon vertex. An inflection point is identified if all vertices under one of the segments have positive curvature and all vertices under the another one have negative curvature. Third algebraic moments (ix) of the region shape provide another way to determine the unknown rotation. Following the method described in [12], the third moments of the shape-normalised region form a complex number $c = \mu'_{x^3} + \mu'_{xy^2} + i(\mu'_{x^2y} + \mu'_{y^3})$, whose phase angle $\alpha = \tan^{-1}(\frac{\mu'_{x^2y} + \mu'_{y^3}}{\mu'_{x^3} + \mu'_{xy^2}})$

changes covariantly with rotation. The last approach used to fix the rotation exploits straight linear segments on the region boundary (viii). A standard Douglas-Peucker algorithm [5, 25] is executed on the shape-normalised region.



(a) (b)

Fig. 5. Example of region concavities. (a) A detected non-convex region with identified concavities and their covariance matrices (b) The largest concavity: the bitangent line and farthest points on the concavity and on the region.

Concavities (v) are identified with segments of the region boundary that depart from the convex hull of the region. For each concavity, two points are found locally maximising distance to the corresponding bi-tangent line (vi). One of them is located on the contour segment that forms the concavity, the other one on the rest of the contour. Figure 5 illustrates a complex, non-convex region with six concavities. Figure 5 (a) shows the centre of gravity and the covariance matrix for each concavity. Figure 5 (b) demonstrates, for one of the concavities, the two points of locally maximal distance.

2.3 LAF Construction

A frame is constructed by combining affine-covariant primitives which, in correspondence, constrain all of the six degrees of freedom. The combinations we used in the experiments are illustrated in Figure 3. The images show basis vectors of the frames along with the primitives – points (e.g. inflection points), linear segments (e.g. bitangents), and ellipses representing covariance matrices. Figure 3 includes a table listing, for each of the frame types, the combination of primitives that define it.

3 Normalisation of Measurement Region

Object recognition from a single training view requires an object representation that does not change (is invariant) if the object is seen from different viewpoints and under different conditions, such as illumination. The previous section detailed constructions of local affine-covariant coordinate systems that are fully defined by image measurements. As such, they "stick" to the objects in the image if the viewpoint changes, and serve as object-centred frames of reference. Invariance of the object representation to geometric variations is thus achieved by normalising local appearance according to the detected frames. Image neighbourhood of every LAF is transformed into a canonical coordinate system, and a geometrically normalised patch is constructed. The patch is then normalised photometrically.

Geometric normalisation. The affine transformation between the canonical frame with origin $O = (0,0)^T$ and basis vectors $e_1 = (1,0)^T$ and $e_2 = (0,1)^T$ and a frame established in the image is described in homogenous coordinates by a 3 by 3 matrix

$$\mathsf{A} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix}.$$

Measurement region (MR) is the part of the image, defined in terms of the affine frame, whose appearance, after appropriate encoding (see Section 4), is used to determine local correspondences. Each local affine frame is associated with one, or possibly multiple, MRs. The choice of MR shape and size is arbitrary. Larger MRs have higher discriminative potential, but are more likely to cover part of an object that is not locally planar. Based on experimental evaluation, our choice is to use a square MR centred around a detected LAF, specifically a region spanning $\langle -2, 3\rangle \times \langle -2, 3\rangle$ in the frame coordinate system. In image coordinate system, the measurement region of a frame A becomes a parallelogram with corners at (in homogenous coordinates):

$$c_1 = \mathsf{A}\begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}, \quad c_2 = \mathsf{A}\begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix}, \quad c_3 = \mathsf{A}\begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}, \quad c_4 = \mathsf{A}\begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix},$$

Photometric Normalisation. A linear camera (i.e. a camera without gamma-correction) is assumed and specular reflections and shadows are ignored. The combined effect of different scene illumination and camera and digitiser settings (gain, shutter speed, aperture) is modelled by affine transformations of individual colour channels, leading to the photometric transformation between two corresponding patches I and I' in the form:

$$\begin{pmatrix} r' \\ g' \\ b' \end{pmatrix} = \begin{pmatrix} m_r & 0 & 0 \\ 0 & m_g & 0 \\ 0 & 0 & m_b \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix} + \begin{pmatrix} n_r \\ n_g \\ n_b \end{pmatrix}$$

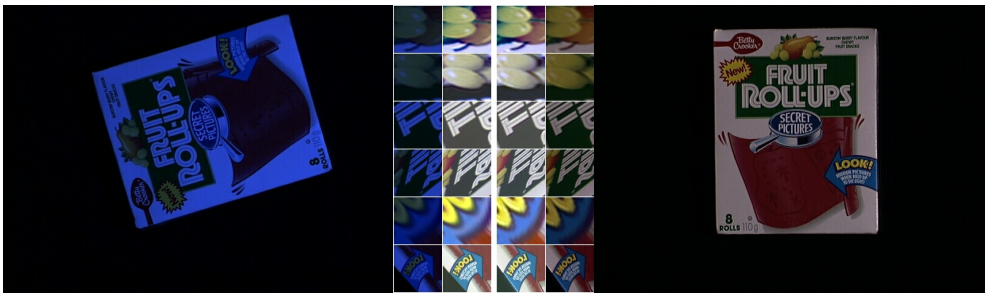
The parameters m_r, n_r, m_g, n_g, m_b, n_b differ for individual correspondences. This model agrees with the monochromatic reflectance model [11] in the case of narrow-band sensor. It can be viewed as an affine extension of the diagonal model that has been shown by Finlayson to be sufficient in common circumstances [7], at least in conjunction with sensor sharpening [8]. To represent a patch invariantly to photometric transformations, intensities are transformed into a canonical form. The intensities of individual colour channels are affinely transformed to have zero mean and unit variance. The normalisation procedure of a local patch is summarised in algorithm 2.

Algorithm 2: Normalisation of a Local Representation

1. Establish a local affine frame, form the affine transformation $\mathbf{A}$ between a canonical coordinate system and the detected system.
2. Express the intensities of the A's measurement region in the canonical coordinate system $I'(\mathbf{x}) = I(\mathsf{A}\mathbf{x}), \quad \mathbf{x} \in$ MR with some discretisation.
3. Apply the photometric normalisation $\hat{I}'(\mathbf{x}) = (I'(\mathbf{x}) - \mu)/\sigma, \quad \mathbf{x} \in$ MR, where μ is the mean and σ is the standard deviation of I' over the MR.

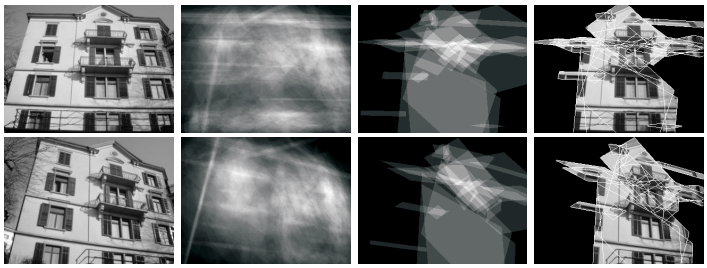
The twelve normalisation parameters ($a_1 \ldots a_6$ for geometric normalisation, m_r, n_r, m_g, n_g, m_b and n_b for photometric normalisation) are stored along with the descriptor of the normalised local patch. When considering a pair of patches for a correspondence, these twelve parameters are combined to provide the local transformations (both geometric and photometric) between the images. The transformations are exploited later during the matching step, as described in Section 4.1.

Figure 6 illustrates the normalisation procedure. On query (a) and database (f) images, MSERs are detected and LAFs constructed, independently on each image. Geometric normalisation according to the transformation between detected LAFs and the canonical coordinate system yields patches depicted in columns (b) and (e). Finally, the result of photometric normalisation of individual patches is shown in columns (c) and (d).



(a) (b) (c) (d) (e) (f)

Fig. 6. Normalised local image patches. (a), (f): Query and Database images, (b), (e): Examples of geometrically normalised MRs (measurement regions), (c), (d): Photometrically normalised MRs



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

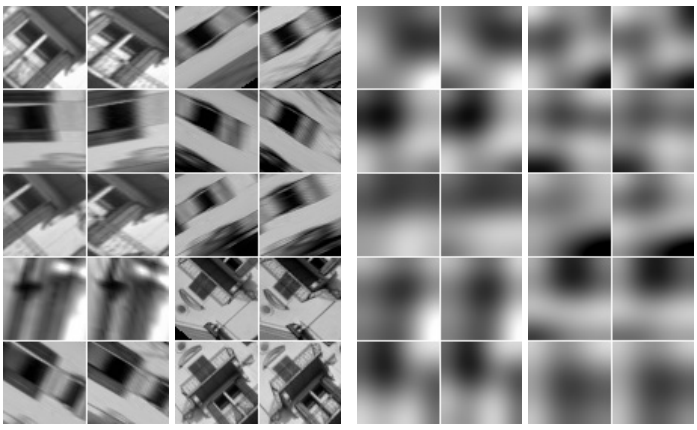
Fig. 7. Example of coverage of images by local patches. (a) original query and database images, (b) image coverage by local patches, whiter area – more overlapping patches, (c) image patches where correspondences between the images were found, (d) image area covered by the corresponding patches.

4 Descriptors of Local Appearance

A descriptor is a suitable data representation of a local image patch. It is associated with a similarity measure, often Euclidean distance. Because of the normalisation, any representation of the normalised patches (shown in Figure 6 (c) and (d)) is theoretically invariant to affine geometric and diagonal photometric transformations. There is therefore no need for e.g. rotation invariance of the representation. Obviously, directly the intensities of the normalised regions can be stored, but such a representation is sensitive to image noise and to imprecise alignment.

The following summarises the desirable properties of a descriptor. A descriptor has to be discriminative, to be able to distinguish between a large number of image regions. The similarity measure should well separate corresponding and not-corresponding regions. The ratio of similarities of matching and mismatched frames (discussed e.g. in Lowe's work [16]) should be maximised. The descriptor should be robust or invariant (i) to localisation errors of the detector, i.e. to misalignment of corresponding representations, and (ii) to image transformations not covered by the detector covariance. If the detector, for example, does not resolve rotation (as various feature point detectors do not) rotational invariants have to be used as descriptors. In our case, local affine frames provide covariance with affine transformations of the image. Our descriptor should thus be insensitive to small perspective distortion and to distortions caused by non-planarity of the surfaces. Finally, the descriptor should be efficient from the computational point of view. The data representation should be compact, to be memory efficient, and fast to construct. More importantly, efficient evaluation of similarity of two descriptors is required.

Discrete Cosine Transformation We represent the local appearance by low-frequency coefficients of the discrete cosine transformation (DCT). For uniformly distributed data, the DCT approximates the Karhunen-Loeve transformation (KLT) [13], which is widely used in pattern recognition to reduce data dimensionality without significant deterioration of recognition rate. DCT has the desirable properties of a descriptor. It is computationally efficient, fast algorithms exist that computes DCT with $O(n \log n)$ time complexity. Hardware implementations of DCT are widely available due to its widespread use in image and video compres-



(a) (b)

Fig. 8. Examples of correspondences established between frames of query (left columns) and database (right columns), for the image pair from Figure 7. (a) geometrically and photometrically normalised image patches, (b) the same patches reconstructed from 10 DCT coefficients per colour channel.

sion (JPEG, MPEG, etc.). Robustness to frame misalignment is achieved by storing only low-frequency coefficients, which are less sensitive to the misalignment than higher frequencies. Discriminativity of the DCT representation depends on the number of coefficients stored. In Section 5, it is experimentally shown how the number of coefficients affect the recognition performance, and that DCT representation outperforms descriptor composed of directly the normalised pixels. Our experiments showed that the DCT representation has about the same discriminative potential as the widely used SIFT descriptor [16].

In Figure 8 (b) an example is shown of what information is preserved if 10 DCT coefficients per colour channel are used. The image patches are the same as in Figure 8 (a).

4.1 Matching, Tentative Correspondences of Local Regions

Let us have a set $\mathcal{S}^D$ of frames F^D detected on a single database image, and a set $\mathcal{S}^Q$ of frames F^Q detected on a query image. Let each frame be associated with a descriptor of normalised local appearance. The set of tentative correspondences $\mathcal{T}$ is a subset of $\mathcal{S}^D \times \mathcal{S}^Q$ where $\times$ denotes the cartesian product. Frame pairs $\{F^D, F^Q\} \in \mathcal{T}$ iff F^D and F^Q are considered potentially corresponding on the basis of local measurements (described later). The correspondences in $\mathcal{T}$ include many outliers as they are based solely on the properties of the two frames in question, regardless of other correspondences on the objects. At a later stage, the correspondences are verified and pruned according to consistency with a global model. Different strategies can be employed to obtain the set $\mathcal{T}$:

Nearest match. This is the most commonly used strategy, used in all the experiments described in Section 5: For each frame $F^Q \in \mathcal{S}^Q$ find closest frame $F^D \in \mathcal{S}^D$: $F^D = \mathrm{argmin}_i(d(F^Q, \mathcal{S}_i^D))$. $\{F^Q, F^D\} \in \mathcal{T}$ iff $d(F^Q, F^D) < \Theta_d$, where d is a "similarity" function discussed later.

Mutually nearest match. This strategy is suitable for symmetric matching problems, e.g. for wide-baseline stereo matching. The fraction of correct correspondences (inliers) in $\mathcal{T}$ is increased, causing the successive global consistency check execute faster. But the absolute number of inliers is typically reduced. For each frame $F^Q \in \mathcal{S}^Q$ find closest frame $F^D \in \mathcal{S}^D$: $F^D = \mathrm{argmin}_i(d(F^Q, \mathcal{S}_i^D))$: For the F^D find closest frame $\overline{F^Q} \in \mathcal{S}^Q$: $\overline{F^Q} = \mathrm{argmin}_i(d(F^D, \mathcal{S}_i^Q))$. $\{F^Q, F^D\} \in \mathcal{T}$ iff $\overline{F^Q} = F^Q \wedge d(F^Q, F^D) < \Theta_d$.

All (or N most) similar. This strategy is used when repetitive structures are expected on the objects of interest. Repetitive structures induce ambiguous correspondences, which cannot be resolved at the time of forming of $\mathcal{T}$. Here, each query frame is associated with a set of possibly corresponding frames – of which at most one is correct. The resolution about which of the correspondences is the correct one (if any) is left to the phase of verification of the global consistency. The drawback is in higher number

of false correspondences (outliers), leading to increase of the computational load of the consistency check, or even to its failure due to small fraction of inliers: For each frame $F^Q \in \mathcal{S}^Q$ find all near frames (or N closest frames) $F^D \in \mathcal{S}^D$. $\{F^Q, \mathcal{S}_i^D\} \in \mathcal{T}$ iff $d(F^Q, \mathcal{S}_i^D) < \Theta_d$.

The function d is a scalar function expressing similarity of two frames. Besides reflecting the similarity of the descriptors of the normalised patches, it might include terms related to the probability of the geometric and photometric transformations between the two frames.

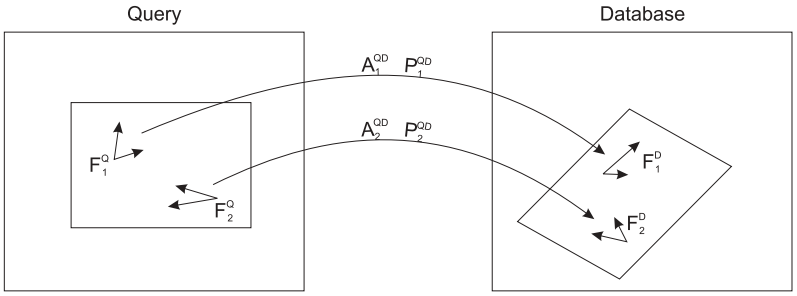


Fig. 9. Illustration of query to model transformations estimated from individual frame correspondences.

Let F^D and F^Q denote the frames on query resp. database images. Let A^D and A^Q be the affine geometric transformations which transform the canonical coordinate system into image coordinates of the respective frames. Finally, let P^D and P^Q be the photometric transformations of the RGB values transforming the normalised intensities to the corresponding intensities in the images. Then the transformations $\mathsf{A}^{QD} = \mathsf{A}^D * (\mathsf{A}^Q)^{-1}$ and $\mathsf{P}^{QD} = \mathsf{P}^D * (\mathsf{P}^Q)^{-1}$ are the geometric resp. photometric transformations between the images – if the frames F^D and F^Q correspond. The situation is illustrated in Figure 9.

Generally, the probability distributions of the transformations A^{QD} and P^{QD} should be estimated from training scenes, and the frame similarity d should be penalised for unlikely transformations. In our experiments the probability distributions are approximated by a step function. If the transformations are out of allowed, problem-specific limits, the frame pair will not match, i.e. d evaluates to infinity. If they are within the limits, no penalty is imposed, and d evaluates directly to the similarity of the descriptors. It allows the function d to be implemented as a fast sequence of thresholdings.

4.2 Globally Consistent Subset of Tentative Correspondences

The process of obtaining tentative correspondences by pair-wise matching of local frames and their descriptors does not take into account the mutual relation between frames. It might for example happen that one of the tentative correspondences implies that the object is larger in the query image than in the model image, while another correspondence suggests that it is smaller and perhaps rotated. Such correspondences, although perfectly possible on their own, are not mutually consistent (assuming the object is rigid). A subset of the obtained tentative correspondences is therefore sought where all correspondences would be consistent with some global object model.

The first issue is the choice of the type of the global model. For general rigid 3D objects the obvious pick is a 3D model imposed through epipolar geometry. A method for estimating epipolar geometry from frame correspondences is described in [3]. The method takes advantage of the fact that a frame correspondence provides an affine transformation between the images, and consequently only three correspondences suffice to obtain the epipolar geometry. For deformable non-rigid (but not articulated) objects, an iterative method described in [6] can be used, although it is rather slow for practical exploitation.

For the purpose of object recognition, simpler models are employed. Unless we are recognising whole complex scenes (e.g. interior of a building), the depth of the visible part of an objects is typically too small to allow for reliable epipolar geometry estimation. We found it sufficient to model the object either as a single planar surface, or as a set of planar surfaces.

Let us have two tentative correspondences, between frames F_1^Q and F_1^D, and between F_2^Q and F_2^D respectively. Each correspondence suggest geometric A_1^{QD} resp. A_2^{QD} and photometric P_1^{QD} resp. P_2^{QD} transformation between the images. Would the frames lie on the same planar surface, the geometric transformations would be identical up to perspective distortion and an imprecision in frame localisation. Assuming light sources at infinity and no shadows nor specular reflections across the planar surface, the two photometric transformations would be also identical.

The set of tentative correspondences $\mathcal{T}$ is decomposed to subsets of consistent correspondences, i.e. subsets in which all correspondences imply identical image-to-image transformation. Each subset represents single plane in the scene. Subsets of low cardinalities are rejected as outliers, and the decision about the presence of an object in the scene relies only on the correspondences in subsets of high cardinality.

5 Experimental validation

The performance of the proposed method was evaluated on several datasets. The COIL-100 dataset has been widely used in object recognition literature [31, 24, 15, 2, 32], and the experiment is included to compare the recognition rate with other state-of-the-art methods. The ZuBuD dataset represents a larger, real-world problem, with images taken outdoor, with occluded objects, varying background, and illumination changes. Finally, FOCUS database represents a retrieval problem, where product logos are sought in scanned advertising material. Typically, the logos occupy only a small portion (e.g. 1%) of the image.

(a) (b)

Fig. 10. COIL-100: (a) Objects from the database, (b) Query images for the occlusion experiment

(a) (b)

Fig. 11. ZuBuD dataset [27]: Examples of (a) query and (b) the corresponding database images.

COIL-100. The Columbia Object Image Library (COIL-100)[1] is a database of colour images of 100 different objects; 72 images of each object placed on a turntable were acquired at pose intervals of 5°. Neither occlusion, background clutter, nor illumination changes are present. Several images from the database are shown in Figure 10(a). Two experiments were performed, differing in the number of images used for training. The achieved recognition rate was 98.2% for 4 training views per object (90° apart, 68 test views per object) and 99.7% for 8 training views (45° apart, 64 test views). Table 2 summarises the results and provides comparison to other published results.

[1] http://www.cs.columbia.edu/CAVE

In another experiment, occlusion of the objects was simulated by blanking one half of the test images (see Figure 10 (b)). Four full (unoccluded) training views per object were used in training. The recognition rate was 87%, which is comparable to published results on unoccluded images. Table 3 provides detailed information about the experiments.

Method	Recognition rate	
	8 training views/object	4 training views/object
MSER+LAF	99.8%	98.2 %
Spectral representation [15]	96.3%	–
Kullback-Leibler SVM [31]	95.2%	84.3%
SNoW / edges [32]	89.2%	88.3%
Spin-Glass MRF [2]	88.2%	69.4%
SNoW / intensity [32]	85.1%	81.5%
Linear SVM [32]	84.8%	78.5%
Nearest Neighbour [32]	79.5%	74.6%

Table 2. COIL-100 experiment: Comparison with published results.

MSER+LAF	COIL-100			ZuBuD
1. Occluded queries	no	no	yes	n/a
2. Training view dist	90°	45°	90°	n/a
3. Number of DB images	400	800	400	1005
4. Number of DB frames	186346	385197	186346	251633
5. Number of query images	6800	6400	6800	115
6. Avg number of query frames	494	494	269	1594
7. avg time to build representation	520 ms	522 ms	251 ms	1255 ms
8. avg recall time	493 ms	3471 ms	277 ms	27234 ms
10. recognition rate	98.24%	99.77%	87.01%	100%

Table 3. Experimental results on COIL-100 and ZuBuD datasets

The ZuBuD dataset. The experiment was conducted on a set of images of 201 buildings in Zurich, Switzerland, which is publicly available [27]. The database consists of five photographs of the 201 buildings. A separate set of 115 query images is provided. For every query image, there are exactly five matching images of the same building in the database. Not all the database buildings have corresponding queries, the number of queries per building ranges from 0 to 5. Query and database images differ in viewpoint; variations in the illumination are present, but rare. Examples of corresponding query and database images are shown in Figure 11.

In the experiment, 115 query images were matched against 1005 database images, ie. 115575 matches were evaluated in total. For every query image, the R closest database images were retrieved. The recall rate r_R was evaluated, which is defined as $r_R = \frac{n_R}{N}$, where n_R is the number of correct answers in the first R retrieved images, and N the number of all possible correct answers. In our case, when every query has 5 corresponding images in the database, $N = \min(R, 5)$.

Two local patch representations (see Sect. 4) are compared, the directly stored intensities versus the DCT coefficients. The results are summarised in Table 4. For both methods, recall r_R is shown for $R = 1 \ldots 5$. The recall r_1 is equivalent to the percentage of correct images retrieved in rank 1. The last column shows the memory required to store the representation of the whole database of 1005 images. The last lines in Table 4 show other results.

The proposed retrieval system performed well, the retrieval performance was, or was close to, 100% in the first rank. The DCT representation performed slightly better than the direct intensity representation, due

to the insensitivity to image noise and small frame misalignments. Regarding the memory requirements, the DCT representation is much more compact. The memory usage is reduced to circa 20–30% depending on the number of DCT coefficients stored.

Method	Average recall r_R					Memory
	r_1	r_2	r_3	r_4	r_5	usage
direct intensity	98.3%	96.6%	93.6%	89.1%	81.9%	1300 MB
DCT 6 coeffs	99.1%	98.3%	95.7%	91.1%	84.0%	290 MB
DCT 10 coeffs	99.1%	98.7%	96.8%	92.2%	85.0%	370 MB
DCT 15 coeffs	100.0%	99.1%	97.4%	92.8%	85.4%	470 MB
HPAT [26]	86.1%					
Random subwindows [17]	95.7%					

Table 4. ZuBuD: Summary of experimental results.

Fig. 12. FOCUS: Query localisation examples. Query images, database images, and query localisations

Fig. 13. FOCUS: Examples of query (left) and database images (right) not retrieved

The FOCUS database contains 360 colour high-resolution images of commercials scanned from miscellaneous magazines. Figure 12 illustrates example queries and identified commercials from the database. For comparison purposes, we run an experiment with an identical setup as the SEDL system introduced by Cohen [4]. The quality of the retrieval is assessed by the same two quantities as defined by Cohen, the recall rate r_R and the precision ρ_R:

$$r_R = \frac{n}{N} \qquad \rho_R = \frac{\sum_{i=1}^{n}(R+1-r_i)}{\sum_{i=1}^{n}(R+1-i)} \tag{4}$$

where n is the number of correct answers in the first R retrieved images, N the number of all correct answers contained in the database, and r_i the rank of the i-th correctly retrieved answer.

SEDL		LAFs	
recall r_{20}	avg precision ρ_{20}	recall r_{20}	avg precision ρ_{20}
70/90 = 77.8%	88%	75/90 = 83.3%	93.5%

Table 5. FOCUS: Retrieval performance compared to the SEDL system.

In Table 5, average recall rate r_{20} and average precision ρ_{20} are given for the number of retrieved images $R = 20$. For each of the 25 queries used by Cohen, the database images were sorted according to the matching score (similarity measure) m, and the recall r_{20} and the precision ρ_{20} were computed according to formula (4). Each of the 25 queries has 2 to 9 correct answers in the database, with the total number of all correct answers equal to 90. The local affine frame (LAF) method achieves a 83% recall, which is approximately 5% better than results reported by Cohen. Note that the LAF method is not attempting to generalise the query (i.e. to categorise). Most database images missed depict *objects different from the query*. Figure 13 shows three such examples. The "failure" in such cases might be viewed as a strength, demonstrating the very high selectivity of the method, distinguishing items that superficially look identical, while being immune to severe affine deformations.

6 Conclusions

An object recognition method representing object appearance by a set of local measurements was described. Invariance to affine transformations is achieved by expressing local appearance in terms of affine-covariantly detected local coordinate systems.

An overview and classification of affine covariant constructions was presented, covariance of the constructions was proven, and computational issues were discussed. The choice of suitable representation of the local appearance, and the problem of formation of tentative region-to-region correspondences were investigated.

It was shown experimentally that the method achieves state-of-the-art results on publicly available object recognition tests (COIL-100, ZuBuD, FOCUS). Change of scale, illumination conditions, out-of-plane rotation, occlusion, locally anisotropic scale change and 3D translation of the viewpoint were all present in the test problems.

Appendix

A Proofs of Affine Covariance of LAF Primitives

Bellow we show that the construction used to establish local affine frames are indeed covariant with affine transformation. In particular, we show how the area, centre of gravity, and covariance matrix of a region changes under affine transformations of the region, and that the properties of tangent points and of the farthest-from-a-line points are maintained.

Area. Consider a region Ω_1, and its transformed image $\Omega_2 = \mathsf{A}\Omega_1$, i.e. $\Omega_2 = \{\mathbf{x}_2 | \mathbf{x}_2 = \mathsf{A}^\mathsf{T}\mathbf{x}_1 + \mathbf{t}; \mathbf{x}_1 \in \Omega_1\}$ The area of Ω_2 is given as

$$|\Omega_2| = \int_{\Omega_2} \mathrm{d}\Omega_2 = \int_{\Omega_1} |\mathsf{A}| \, \mathrm{d}\Omega_1 = |\mathsf{A}||\Omega_1|, \tag{5}$$

where $|\mathsf{A}|$ is the determinant of A, and $|\Omega|$ is the area of region Ω. The area of a transformed region equals $|\mathsf{A}|$ times the area of the original region.

Centre of gravity. The centre of gravity of region Ω is $\mu = \frac{1}{|\Omega|}\int_\Omega \mathbf{x}\mathrm{d}\Omega$. The relation between the centres of gravity of transformed regions is:

$$\begin{aligned}\mu_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2}\mathbf{x_2}\,\mathrm{d}\Omega_2 = \frac{1}{|\mathsf{A}||\Omega_1|}\int_{\Omega_1}(\mathsf{A}^\mathsf{T}\mathbf{x_1}+\mathbf{t})|\mathsf{A}|\,\mathrm{d}\Omega_1 = \mathsf{A}^\mathsf{T}\frac{1}{|\Omega_1|}\int_{\Omega_1}\mathbf{x_1}\,\mathrm{d}\Omega_1 + \frac{1}{|\Omega_1|}\int_{\Omega_1}\mathbf{t}\,\mathrm{d}\Omega_1 \\ &= \mathsf{A}^\mathsf{T}\mu_1 + \mathbf{t}, \end{aligned} \tag{6}$$

the centre of gravity changes covariantly with the affine transform.

Covariance matrix. The covariance matrix Σ of a region Ω is a 2x2 matrix defined as $\Sigma = \frac{1}{|\Omega|}\int_\Omega(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\mathsf{T}\,\mathrm{d}\Omega$. Covariance matrix of a transformed region Ω_2 is then

$$\begin{aligned}\Sigma_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2}(\mathbf{x_2}-\mu_2)(\mathbf{x_2}-\mu_2)^\mathsf{T}\,\mathrm{d}\Omega_2 \\ &= \frac{1}{|\mathsf{A}||\Omega_1|}\int_{\Omega_1}(\mathsf{A}^\mathsf{T}\mathbf{x_1}+\mathbf{t}-(\mathsf{A}^\mathsf{T}\mu_1+\mathbf{t}))(\mathsf{A}^\mathsf{T}\mathbf{x_1}+\mathbf{t}-(\mathsf{A}^\mathsf{T}\mu_1+\mathbf{t}))^\mathsf{T}|\mathsf{A}|\,\mathrm{d}\Omega_1 \\ &= \frac{1}{|\Omega_1|}\int_{\Omega_1}(\mathsf{A}^\mathsf{T}(\mathbf{x_1}-\mu_1))(\mathsf{A}^\mathsf{T}(\mathbf{x_1}-\mu_1))^\mathsf{T}\,\mathrm{d}\Omega_1 = \mathsf{A}^\mathsf{T}\left(\frac{1}{|\Omega_1|}\int_{\Omega_1}(\mathbf{x_1}-\mu_1)(\mathbf{x_1}-\mu_1)^\mathsf{T}\,\mathrm{d}\Omega_1\right)\mathsf{A} \\ &= \mathsf{A}^\mathsf{T}\Sigma_1\mathsf{A} \end{aligned} \tag{7}$$

Cholesky decomposition of a symmetric and positive-definite matrix Σ is a factorisation $\Sigma = U^\mathsf{T}U$, where U is an upper triangular matrix. Cholesky decomposition is defined up to a rotation, since $U^\mathsf{T}U = U^\mathsf{T}R^\mathsf{T}RU$ for any orthonormal R. For the decomposition of covariance matrix of a transformed region we write

$$\Sigma_2 = U_2^\mathsf{T}R_2^\mathsf{T}R_2U_2 = \mathsf{A}^\mathsf{T}U_1^\mathsf{T}R_1^\mathsf{T}R_1U_1\mathsf{A} = \mathsf{A}^\mathsf{T}\Sigma_1\mathsf{A}, \qquad \text{thus} \qquad U_2^\mathsf{T} = \mathsf{A}^\mathsf{T}U_1^\mathsf{T}R \tag{8}$$

Hence the triangular matrix U obtained as the Cholesky-decomposition of a covariance matrix Σ is covariant, up to an arbitrary orthonormal matrix R, with the affine transform applied to the region.

Line parallelism. Let us consider two lines, determined by points $\mathbf{p}$ and $\mathbf{q}$, and $\mathbf{r}$ and $\mathbf{s}$ respectively. The lines are parallel, iff

$$(\mathbf{p}-\mathbf{q}) = k(\mathbf{r}-\mathbf{s}),\ k\in\mathbb{R}\setminus\{0\}$$

Affinely transformed lines are then parallel iff

$$\begin{aligned}(\mathsf{A}^\mathsf{T}\mathbf{p}+\mathbf{t}-\mathsf{A}^\mathsf{T}\mathbf{q}-\mathbf{t}) &= k(\mathsf{A}^\mathsf{T}\mathbf{r}+\mathbf{t}-\mathsf{A}^\mathsf{T}\mathbf{s}-\mathbf{t}) \\ \mathsf{A}^\mathsf{T}(\mathbf{p}-\mathbf{q}) &= k\mathsf{A}^\mathsf{T}(\mathbf{r}-\mathbf{s}) \\ (\mathbf{p}-\mathbf{q}) &= k(\mathbf{r}-\mathbf{s}) \end{aligned} \tag{9}$$

which is true if and only if the lines were parallel before the transformation. Thus, affine transformation preserves line parallelism.

Ordering of distances to a line: Let us have a line determined by two points $\mathbf{p}$ and $\mathbf{q}$. For a point $\mathbf{x}$, its distance d_1 to the line $\mathbf{pq}$ is $d_1 = \frac{2S}{|\mathbf{p}-\mathbf{q}|}$, where S is the area of the $\mathbf{pqx}$ triangle. Using eq. 5, it follows that the transformed distance d_2 is given by

$$d_2 = \frac{2|\mathsf{A}|S}{|A^\mathsf{T}\mathbf{p}+\mathbf{t}-\mathsf{A}^\mathsf{T}\mathbf{q}-\mathbf{t}|} = \frac{|\mathsf{A}||\mathbf{p}-\mathbf{q}|}{|\mathsf{A}^\mathsf{T}\mathbf{p}-A^\mathsf{T}\mathbf{q}|}d_1 = kd_1$$

where k is a scalar constant for given line $\mathbf{pq}$ and transformation A. Affine transformation thus preserves ordering of distances of points from a line. It directly follows that a point $\mathbf{x}\in X$ with the property of being of all the points in X the one farthest from a line $\mathbf{pq}$, retains the property under affine transformations.

Incidence of points and lines: Under affine transformations, points incident with a line will remain on the line, and, vice-versa, distinct points will not be brought to the line unless the transformation is singular. The property is again easily shown exploiting the covariance of region area, from Equation. 5. Considering a line defined by two distinct points $\mathbf{p}$ and $\mathbf{q}$, and a point $\mathbf{x}$, the area S_1 of the $\mathbf{pqx}$ triangle equals to zero if $\mathbf{x}$ is on $\mathbf{pq}$ and nonzero otherwise. After affine transformation, the area of the triangle becomes

$S_2 = |A|S_1$, where $|A|$ is the determinant of the transformation matrix (S_2 is the area of triangle given by points defining the transformed line, i.e. $\mathsf{A}^\mathsf{T}\mathbf{p} + \mathbf{t}$ and $\mathsf{A}^\mathsf{T}\mathbf{q} + \mathbf{t}$, and the transformed point $\mathsf{A}^\mathsf{T}\mathbf{x} + \mathbf{t}$). Assuming nonsingular transformation, i.e. $|\mathsf{A}| \neq 0$, the transformed triangle has area $S_2 = 0$ if and only if $S_1 = 0$. Thus the incidence is maintained.

Tangent and bitangent lines: Tangent line is a line incident with region boundary (in a tangent point $\mathbf{p}$), which does not pass through any of the region interior points. Since the incidence property between the tangent line and the boundary, respective interior points, is maintained, the line transformed by an affine transformation remains tangent to the transformed region, and the tangency occur in the transformed point $\mathbf{p}_2 = \mathsf{A}^\mathsf{T}\mathbf{p} + \mathbf{t}$. An analogy holds for the bitangent lines, where both tangent points are maintained.

An affine transformation is either orientation-preserving or orientation-reversing, if determinant $|A|$ is positive or negative respectively [23]. Therefore the sign of the curvature $\kappa = \frac{\mathrm{d}\Phi}{\mathrm{d}s}$ of a transformed region is either reversed or preserved. It follows that **linear segments** of the contour (segments of zero curvature) and **inflection points** (points where the curvature changes its sign, without specifying whether from positive to negative or vice versa) are maintained.

References

1. C. Ballester and M. Gonzalez. Affine invariant texture segmentation and shape from texture by variational methods. *Journal of Mathematical Imaging and Vision*, 9:141–171, 1998.
2. 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.
3. O. Chum, J. Matas, and Š. Obdržálek. Enhancing RANSAC by generalized model optimization. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, volume 2, pages 812–817, January 2004.
4. S. Cohen. Finding color and shape patterns in images. Technical Report STAN-CS-TR-99-1620, Stanford University, May 1999.
5. D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, 10:112–122, 1973.
6. V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *Proceedings of the European Conference on Computer Vision*, volume I, pages 40–54, May 2004.
7. G. Finlayson, M. Drew, and B.Funt. Color constancy: Generalized diagonal transforms suffice. *Journal of the Optical Society of America*, 11:3011–3019, 1994.
8. G. Finlayson, M. Drew, and B. Funt. Spectral sharpening: Sensor transformations for improved color constancy. *Journal of the Optical Society of America*, 11:1553–1563, 1994.
9. P.-E. Forssén and G. Granlund. Robust multi-scale extraction of blob features. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pages 11–18, 2003.
10. C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.
11. G. Healey. Using color for geometry-insensitive segmentation. *Journal of the Optical Society of America*, 6:86–103, June 1989.
12. J. Heikkilä. Pattern matching with affine moment descriptors. *Pattern Recognition*, 37(9):1825–1834, 2004.
13. A. K. Jain. *Fundamentals of Digital Image Processing*. 1986.
14. T. Lindeberg. Feature detection with automatic scale selection. *International Journal on Computer Vision*, 30(2):79–116, 1998.
15. X. Liu and A. Srivastava. A spectral representation for appearance-based classification and recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 37–40, 2002.
16. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 20(2):91–110, 2004.
17. R. Marée, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2005.
18. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.
19. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the International Conference on Computer Vision*, pages 525–531, 2001.
20. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the European Conference on Computer Vision*, pages 128–142, 2002.

21. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(7):43 – 72, November 2005.
22. F. Mokhtarian and A. K. Mackworth. A theory of multiscale, curvature-based shape representation for planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):789–805, 1992.
23. J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. 1992.
24. Š. Obdržálek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proceedings of the British Machine Vision Conference*, 2002.
25. U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1:244–259, 1972.
26. H. Shao, T. Svoboda, T. Tuytelaars, and L. Van Gool. HPAT indexing for fast object/scene recognition based on local appearance. In *International Conference on Image and Video Retrieval*, pages 71–80, 2003.
27. H. Shao, T. Svoboda, and L. Van Gool. ZuBuD — Zurich Buildings Database for Image Based Recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology, March 2003. http://www.vision.ee.ethz.ch/showroom/zubud.
28. J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, pages 1470–1477, 2003.
29. T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Visual Information and Information Systems*, pages 493–500, 1999.
30. T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proceedings of the British Machine Vision Conference*, 2000.
31. N. Vasconcelos, P. Ho, and P. J. Moreno. The Kullback-Leibler kernel as a framework for discriminant and localized representations for visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 430–441, 2004.
32. M. H. Yang, D. Roth, and N. Ahuja. Learning to Recognize 3D Objects with SNoW. In *Proceedings of the European Conference on Computer Vision*, pages 439–454, 2000.