Cognitive Vision Workshop (CogVis02)
September 2002, Zurich, Switzerland

Learning Parameters of a Recognition System Based on Local Affine Frames

Jiří Matas[1,2] and Štěpán Obdržálek[1,2]
[1]Center for Machine Perception, Czech Technical University, Prague, 120 35, CZ
[2]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

Abstract

An approach to object recognition, based on matching of local image features, is presented. First, distinguished regions of data-dependent shape are robustly detected. On these regions, local affine frames are established using several affine invariant constructions. Direct comparison of photometrically normalised colour intensities in local, geometrically aligned frames results in a matching scheme that is invariant to piecewise-affine image deformations, but still remains very discriminative.

Nevertheless, invariance to a wide range of local geometric and photometric transformations reduces the discriminative power – not all possible transformations are equiprobable. Probability of the transformations is estimated from matches established by the invariant method on the training data. The estimate is exploited in the recognition phase to favour local correspondences with more likely transformations.

The potential of the approach is experimentally verified on COIL-100 – a publicly available image database. 99.9% recognition rate is obtained for 18 training views per object.

1 Introduction

Our interest is in an object recognition system which is able to recognise 2D and 3D objects using a representation learned from a given training set of labelled images of the objects. A recognition system, plausibly aspiring to at least partially emulate the capabilities of humans, shall have a number of desirable attributes. It should

- be able to learn the object representation from only a few examples.
- be stable under viewpoint and illumination changes.
- be robust to occlusion and background clutter.
- be able to deal with a large number of objects.
- support fast indexing for at least coarse decisions (reducing the number models considered in detail).
- be able to exploit small differences for distinguishing between similar objects.

Towards meeting these demands, we proposed an approach that is able to recognise hundreds of 3D objects, achieving high recognition rates (over 99% for 8 training views and over 95% for four training views on the COIL-100 database [1], 100% recognition rate from a single training view on the SOIL-47 database [2]). In our previous work, a representation based on *local affine-invariant descriptors* was learned from examples [15]. In the work presented here, the system is modified to learn the probability distribution of the parameters of the local photometric and geometric transformation.

Approaches to Object Recognition. In general, two main trends can be distinguished: model-based and appearance-based approaches. While model-based methods try to analytically model the relation between the object and its projection to the image, appearance-based methods recognise objects by visual similarity. Model-based approaches usually rely on extraction of 2D primitives, such as image edges, which are hard to obtain and interpret reliably. On the other hand, appearance-based approaches, that directly use the intensity function or transformation thereof (eigenimages, colour histograms, etc.), are prone to fail under viewpoint and illumination changes, once the appearance of the object changes substantially.

As an attempt to combine advantages of both approaches, methods based on the matching of local features have been proposed. Like in the appearance-based approaches, an object model is learnt from images thereof. However, local features are extracted and used for the matching. The advantage here is that the deformations of object appearance caused by viewpoint changes, although being globally complex, can be approximated by simple transformations at the local scale. Various methods in this category differ in the choice of local image regions and in the features computed over these regions [5, 9, 12, 4, 10, 7, 14, 13].

Invariance in object recognition. In general, invariance to a more complex transformation means better immunity to changes in the image formation process, but also reduction in discriminative power. One would like to be invariant to only such changes, that are really going to happen. Let us consider an example. Arbitrarily placed objects are to be recognised in an indoor environment, so that while we need an invariance to geometric deformations caused by viewpoint changes, the illumination conditions (spectral power distribution) can be considered constant. There is no need to be photometrically invariant to an affine transformation of colours or even a diagonal transformation. Assuming that the colour vector undergoes a simple scaling (scalar multiplication) is sufficient and covers many effects from the change in illumination intensity, change in aperture or exposure time to effects due to the change of orientation w.r.t. the light source (under the monochromatic model, ignoring specularities).

Even invariance to such a weak model of photometric change reduces the discriminative power of recognition system. White matches any grey or even black equally well, but in practice, such dramatic changes of the multiplicative factor very rarely happen. In recognition, it is valuable to know the *possible* transformation of measurements, but it is equally important to exploit what is *probable*.

Systems exploiting such constraints would clearly be superior in recognition performance to systems aimed at general illumination conditions. The a priori knowledge about the constraints can be hard-wired into the system at the design time, but this leads to an ad hoc system fine-tuned to specific conditions. Alternatively, a recognition system should start in unknown conditions as a system based on invariants, gradually learning automatically the probabilities of various transformation parameters and improving its performance.

In a general framework, we may never be sure that the probabilities are stationary (e.g. when the system is moved to another room) and such situations must be taken care of. Thus a recognition system can operate as a mixture of systems for known and unknown conditions. The probability of observing an object O for a given measurement X is then

$$\begin{aligned} P(O|X) &= P(s)P(O|D(X)) \\ &+ (1-P(s))P(O|I(X)) \end{aligned} \quad (1)$$

where $P(s)$ is the probability that the learned transformation probabilities are still valid, $D()$ is a discriminative and non-invariant description, and $I()$ is an invariant description of the measurements. The probabilities should be ideally updated online.

In summary, learning may reduce the need for invariance. More extensive training data allow for better modelling of possible changes in object appearance.

Overview of our approach. We assume that view-dependent image deformations can be reasonably well approximated by local affine transformations of both geometry and illumination. Such assumption holds for objects where locally planar surface regions can be found, and where the size of such regions is small relative to the camera distance, so that perspective distortions can be neglected. The proposed approach is based on a robust, affine invariant detection of local affine frames (local coordinate systems). Local correspondences are established by a direct comparison of normalised colours in image patches represented canonically in normalised affine frames.

The method is thus in general invariant to local affine transformations of both geometry and photometry. In this paper, first experiments are carried out towards a system learning the statistics of the photometric and geometric transformations. The long-term objective is to create a mixed system that, when appropriate, would improve its recognition performance by exploiting statistics about photometric and geometric transformations estimated on the training set.

The paper is organised as follows. In Section 2 we briefly review the concept of distinguished regions. Section 3 gives a description of procedures for construction of local affine frames on the distinguished regions. Section 4 details how local correspondences are established, exploiting the statistics learned on the training dataset. In Section 5 experimental results are presented.

2 Distinguished Regions

Distinguished Regions (DRs) are image elements (subsets of image pixels), that posses some distinguishing, singular property that allows their repeated and stable detection over a range of image formation conditions. In this work we exploit a new type of distinguished regions introduced in [6], the *Maximally Stable Extremal Regions* (MSERs). An extremal region is a connected component of pixels which are all brighter (MSER+) or darker (MSER-) than all the pixels on the region's boundary. This type of distinguished regions has a number of attractive properties: 1. invariance to affine and perspective transforms, 2. invariance to monotonic transformation of image 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. We do not describe the MSERs here; the reader is referred to [6] which includes a formal definition of the MSERs and a detailed description of the extraction algorithm. The report [6] is available online. Examples of detected MSERs are shown in Figure 1. Note that DRs do not form segmentation, since DRs do not cover entire image area, and DRs can be (and usually are) nested.

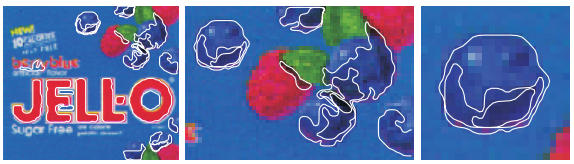


Figure 1. An example of detected distinguished regions of MSER type

3 Local Frames of Reference

Local affine frames facilitate normalisation of image patches into a canonical frame and enable direct comparison of photometrically normalised intensity values, eliminating the need for invariants. It might not be possible to construct local affine frames for every distinguished region. Indeed, no dominant direction is defined for elliptical regions, since they may be viewed as affine transformations of circles, which are completely isotropic. On the other hand, for some distinguished regions of a complex shape, multiple local frames can be affine-invariantly constructed in a stable and thus repeatable way. Robustness of our approach is thus achieved by selecting only stable frames and employing multiple processes for frame computation.

Definition of terms:

Affine transformation is a map $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form $F(\mathbf{x}) = A^T\mathbf{x} + \mathbf{t}$, for all $\mathbf{x} \in \mathbb{R}^n$, where A is a linear transformation of $\mathbb{R}^n$, assumed non-singular here.

Center of gravity (CG) of a region Ω is $\mu = \frac{1}{|\Omega|}\int_\Omega \mathbf{x} d\Omega$.

Covariance matrix of a region Ω is a $n \times n$ matrix defined as $\Sigma = \frac{1}{|\Omega|}\int_\Omega (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T d\Omega$.

Bi-tangent is a line segment bridging a concavity, i.e. its endpoints are both on the region's outer boundary and the convex hull, all other points are part of the convex hull.

Affine covariance of the center of gravity and of the covariance matrix is shown in Appendix A. The invariance of the bi-tangents is a consequence of the affine invariance (and even projective invariance) of the convex hull construction [11, 8]. Finally, we exploit the affine invariance of the maximal-distance-from-a-line property, which is easily appreciated taking into account that affine transform maintains parallelism of lines and their ordering.

A two-dimensional affine transformation possesses six degrees of freedom. Thus, to determine an affine transformation, six independent constraints are to be applied. Various constructions can be utilised to obtain these constraints. In particular, we use a direction (providing a single constraint), a 2D position (providing two constraints), and a covariance matrix of a 2D shape (providing three constraints).

Frame constructions. Two main groups of affine-invariant constructions are proposed, based on 1. region normalisation by the covariance matrix and the center of gravity, and 2. detection of stable bi-tangents

Transformation by the square root of inverse of the covariance matrix normalises the DR up to an unknown rotation. To complete an affine frame, a direction is needed to resolve the rotation ambiguity. The following directions are used: 1. Center of gravity (CG) to a contour point of extremal (either minimal or maximal) distance from the CG 2. CG to a contour point of maximal convex or concave curvature, 3. CG of the region to CG of a concavity, 4. direction of a bi-tangent of a region's concavity.

In frame constructions derived from the bi-tangents, the two tangent points are combined with a third point to complete an affine frame. As the third point, either 1. the center of gravity of the distinguished region, 2. the center of gravity of the concavity, 3. the point of the distinguished region most distant from the bi-tangent, or 4. the point of the concavity most distant from the bi-tangent is used. Another type of frame construction is obtained by combining covariance matrix of a concavity, CG of the concavity and the bi-tangent's direction.

Frame constructions involving the center of gravity or the covariance matrix of a DR rely on the correct detection of the DR in its entirety, while constructions based solely on properties of the concavities depend only on a correct detection of the part of the DR containing the concavity.

Figure 2 visualise the process of shape-normalisation and a dominant point selection. A distinguished region detected in an image is transformed to the shape-normalised frame, the transformation being given by the square root of inverse of the covariance matrix. Normalised contour curvatures and normalised contour distances are searched for stable extremal values to resolve the rotation ambiguity. One of the constructed frames is shown on the right in Figure 2, represented by the two basis vectors of the local coordinate system. Figure 3 shows three examples of the local affine frame constructions based on concavities.

4 Matching

Once local affine frames are computed in a pair of images, (geometrically) invariant descriptors of local appearance are not needed for the matching. Correspondences are established simply by correlating photometrically normalised image intensities in geometrically normalised measurement regions.

Measurement regions (MRs) are defined in local coordinate systems of the affine frames, but the choice about MR

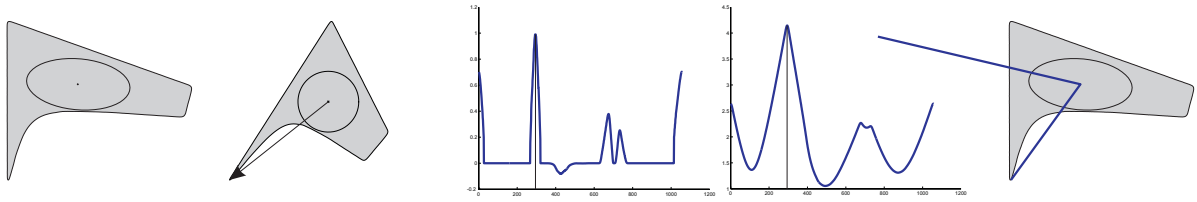
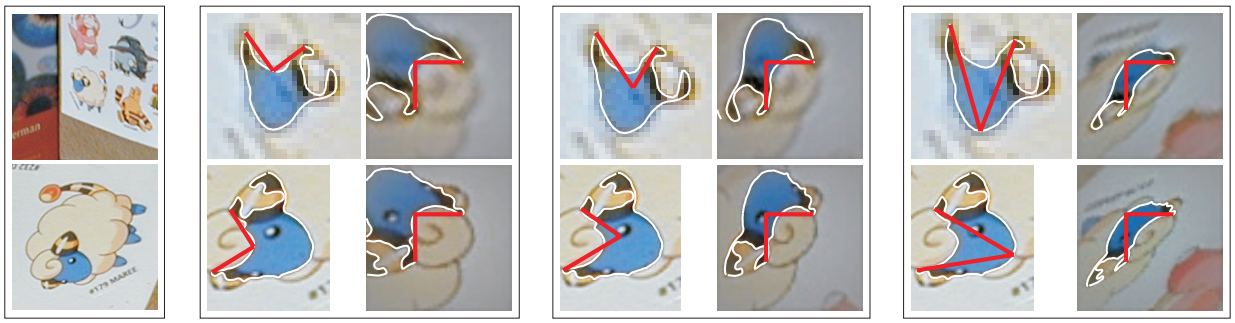


Figure 2. Construction of affine frames. From left to right: a distinguished region (the grey area), the DR shape-normalised according to the covariance matrix, normalised contour curvatures, normalised contour distances to the center of DR, and one of the constructed frames represented by its basis vectors.



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

Figure 3. Bi-tangent based constructions of affine frames. (a) original views, (b) 2 tangent points + farthest concavity point, (c) 2 tangent points + DR's center of gravity, (d) 2 tangent points + farthest DR point. Left columns - detected frames, right columns - locally normalised images

shape and size can be arbitrary. Larger MRs have higher discriminative potential, but are more likely to cover an object area that violates the local planarity assumption. Our choice is to use a square MR centred around a detected LAF, specifically an image area spanning $\langle -2, 3\rangle \times \langle -2, 3\rangle$ in the frame coordinate system. Multiple MRs for every LAF could be used, increasing the robustness (and computational complexity) of the method. The frame normalisation proceeds in four steps:

1. establish a local affine frame
2. compute the affine transformation mapping the LAF to a normalised coordinate system
3. resample the intensities of the LAF's measurement region into a raster in the normalised coordinate system. To represent the content of normalised MRs, we use rasters of size 21×21 pixels.
4. The photometric normalisation $\hat{I}(x, y) = (I(x, y) - \mu)/\sigma, \quad x, y \in \{1, \ldots, 21\}$ is applied, where μ is the mean and σ is the standard deviation of I over MR.

See Figure 3 for examples of frame normalisations.

The twelve normalisation parameters (6 for geometric and 3×2 for photometric normalisations) are stored along with $\hat{I}$. When considering a pair of frames for a correspondence, these parameters are combined to provide a local, between-frame transformation. The correspondences are established by correlating intensities in the normalised frames, weighted by the probability of the local transformation. Figure 4 shows an example of correspondences found for a pair of images from the COIL-100 database.

4.1 Learning the Transformation Probabilities

If the training views are taken densely enough to provide correspondences in between them, we can learn the statistics of the local transformations (both geometric and photometric). The transformation probability is then used to influence the process of determining local correspondences in the recognition phase. Since estimating the distribution in a 12-dimensional space would require enormous amount of training correspondences, we decompose the transformation to a set of low-dimensional transformations. We are

thus considering the individual transformation components as uncorrelated.

Decomposition of Geometric Transformation.

Dropping the translation, a 2D affine transformation A becomes a linear transformation T, which can be represented as a 2×2 matrix (assuming nonsingular). The transformation can be decomposed into

$$T = R_1 D R_2$$

where D is a diagonal matrix of singular values of T, and R_1 and R_2 are unitary orthogonal matrices, ie. rotations combined with a potential mirroring. Further, we decompose the matrix D into an areal scale $s = det(D)$, and an anisotropic scale $a = d_1/d_2$, where d_1 is the greater and d_2 the lesser of the singular values.

If training and test images are of the same nature, the model→image transformation T is equally probable as the image→model transformation T^{-1}, $T^{-1} = R_2^{-1} D^{-1} R_1^{-1}$. This gives us $P(s) = P(1/s)$, $P(a) = P(1/a)$, $P(\phi_1) = P(-\phi_2)$, and $P(\phi_2) = P(-\phi_1)$, where ϕ_1 and ϕ_2 are the angles of the rotations R_1 and R_2 respectively. Estimating only the absolute value of the angles, we get $P(|\phi_1|) = P(|\phi_2|) = P(|\phi|)$, thus reducing the dimensionality from four dimensions to three. Taking the logarithm of the scale s, we get a symmetric distribution $P(log(s)) = P(-log(s))$, which allows us to estimate it as $P(|log(s)|)$. Similarly, the distribution of the anisotropic scale a is transformed to a distribution of $P(log(a))$ (here the logarithm is never negative).

To summarise, we model the distribution of the linear transformation T with a set of three independent one-dimensional random variables, $|log(s)|$, $log(a)$ and $|\phi|$. For examples of the distributions estimated on the COIL-100 database see Figure 5.

Decomposition of Photometric Transformation.

We consider the local photometric transformation to be in form $\overline{x}' = C\overline{x} + \overline{b}$, where $\overline{x} = [R, G, B]^T$, C is a diagonal matrix $C = diag[c_R, c_G, c_B]$ and $\overline{b} = [b_R, b_G, b_B]^T$. We model the change in contrast using a single value u, combining all the colour channels, $u = \sqrt[3]{c_R c_G c_B}$. The matrix C then becomes $C = diag[u, u, u] diag[c'_R, c'_G, c'_B]$. Again, since $P(u) = P(1/u)$ is assumed, $P(|log(u)|)$is estimated instead. Furthermore, b_R, b_G and b_B are considered uncorrelated and equiprobable, so $P(b_R) = P(b_G) = P(b_B) = P(b)$. Currently, only distributions of $|log(u)|$ and $|b|$ are estimated. Figure 5 shows an example of the distributions.

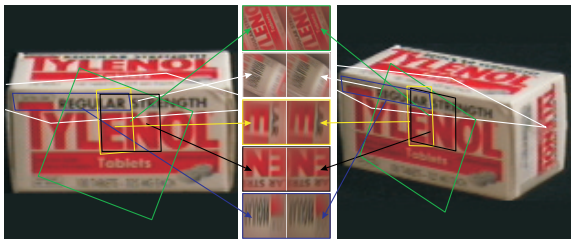


Figure 4. Examples of correspondences established between frames of a training image (left) and a test image (right).

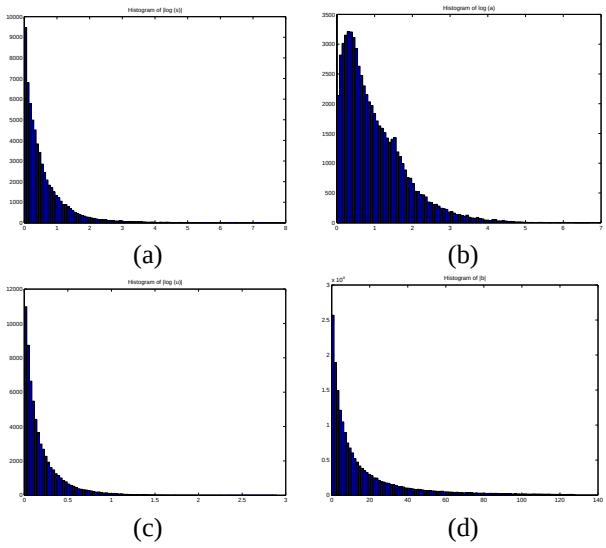


Figure 5. Example of histograms of distributions of transformation parameters. (a) scale $|log(s)|$, (b) anisotropic scale $log(a)$, (c) change in contrast $|log(u)|$, (d) intensity offset $|b|$.

4.2 The Matching Score.

The choice of the best strategy for the computation of the total matching score from individual local correspondences depends on the application. Possible strategies generally differ in the emphasis put on the global model consistency. An extreme approach, used in experiments in this paper, is to ignore the global consistency at all. Counting the number of established local correspondences gives a reasonable estimate of the object similarity; the higher the number of similar local features, the higher the matching score. On the COIL-100 database, this strategy works well when images of the same object viewed from very different viewing angles (up to $180°$) are matched.

The opposite approach is applicable when the model im-

ages are segmented and known to be planar, as may be the case when recognising trademarks, logos, billboards or traffic signs. The model appears in the unknown scene (test image) only as an perspective deformation of the training image. Matching score can be then estimated by maximising the correlation between the whole segmented model and the test image; the set of transformations considered is obtained from local frame correspondences. Other approaches may exploit deformable models, or epipolar geometry constraint for rigid 3D objects.

Matching using Principal Component Analysis. Establishing correspondences by correlating rasters of normalised measurement regions is computationally expensive. We use the principal component analysis (PCA) of the rasters computed over the training set to reduce the dimensionality of the originally $21 \times 21 \times 3$ dimensional data.

Table 1 demonstrates the dependency of the recognition rate on the number of principal components retained. The numbers were obtained on the COIL-100 database, using four training views per object, ie. 400 training images in total. The number of frames was about 100000.

PCA dimension	Recog. rate	% of total variance
5	89.1%	45%
10	91.3%	56%
20	92.7%	67%
50	94.3%	80%
100	94.9%	88%
correlation	95.2%	100%

Table 1. COIL-100: matching frames using PCA, for 4 training views per object

5 Experiments on the COIL-100 database

COIL-100. The Columbia Object Image Library (COIL-100) [1] is a database of colour images of 100 different objects, where 72 images of each object were taken at pose intervals of 5°. The images were preprocessed so that either the object's width or height (whatever is larger) fits the image size of 128 pixels. The COIL-100 (or more often its subset COIL-20) has been widely used in object recognition experiments. In Figure 6 several objects from the database are shown.

Table 2 compares the achieved recognition rates with other 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 insensitivity to image deformations. The LAF approach performs best in all experiments, regardless of the number of training views. For only four training views (90° apart), the recognition rate is almost 95%, demonstrating the remarkable robustness to local affine distortions. In the case of 18 training views per object, only 5 out of the total 5400 test images were misclassified. Table 3 summarises achieved recognition rates up to rank 4.

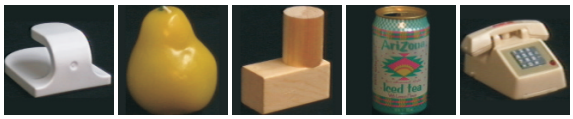


Figure 6. Several objects from COIL-100 database

The results shown here were obtained using thresholding to restrict the range of possible local transformations. No penalty was added to the matching score if all of the transformation components (scale, anisotropic scale, contrast and intensity changes) were in between a respective pair of thresholds. Otherwise, when at least one of the components was outside its thresholds, the respective correspondence was rejected. The thresholds were set manually.

In order to demonstrate the worth of the learning, we have performed an experiment where the matching score was penalised by the transformation improbability. The problem here is that to estimate the probability distributions, local correspondences are to be computed on the training set. When the training views are 90° or more apart, no correspondences between them could be possibly found. We have therefore performed an experiment with 8 training view per object, achieving a recognition rate of 99.3%. The manual setting of thresholds was replaced by an automatic estimate of transformation probabilities at only a minimal decrease in the recognition performance.

Note that we were not building any kind of multi-view object model. If more than one view per object was available for the training, these views were treated independently, as if of different objects. In [15] experimental results on SOIL-47, another publicly available database, are presented.

	# of training views per object				
Rank	18	8	4	2	1
$= 1$	99.9%	99.4%	95.3%	87.8%	76.0%
≤ 2	99.9%	99.7%	96.8%	92.2%	83.2%
≤ 3	99.9%	99.7%	97.3%	95.0%	86.9%
≤ 4	99.9%	99.7%	97.7%	96.2%	89.3%

Table 3. COIL-100: Recognition rate, ranks 1 to 4

training views per object	18	8	4	2	1
total test views	5400	6400	6800	7000	7100
LAFs	99.9%	99.4%	94.7%	87.8%	76.0%
SNoW / edges [16]	94.1%	89.2%	88.3%	-	-
SNoW / intensity [16]	92.3%	85.1%	81.5%	-	-
Linear SVM [16]	91.3%	84.8%	78.5%	-	-
Spin-Glass MRF [3]	96.8%	88.2%	69.4%	57.6%	49.9%
Nearest Neighbour [16]	87.5%	79.5%	74.6%	-	-

Table 2. COIL-100: Recognition rate (rank 1), in comparison to other methods

Occlusions on the COIL-100. We have simulated occlusion of the objects by erasing one half of the test images. The system was trained using full images, again with five different numbers of training views. Figure 7 shows examples of the occluded test images. Recognition rates are summarised in Table 4.

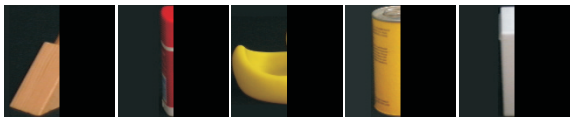


Figure 7. COIL-100: Examples of test images for the occlusion experiment

6 Conclusions

In this paper, an approach to appearance based object recognition was presented. Local affine frames were obtained on distinguished regions of a data-dependent shape, and direct comparison of geometrically and photometrically normalised image patches allowed to establish robust and discriminative local correspondences. Selective matching at the level of local features enabled successful recognition of many objects even when the objects were seen from angles differing by 180° from the training view. Exploiting the principal component analysis, the otherwise high memory and computational costs of the method were reduced.

The probabilities of local transformations (both geometric and photometric) were estimated from matches obtained on the training set. In successful but still preliminary experiments, correspondences were rated by the similarity of the feature vectors as well as by the likelihood of the local transformations induced.

Successful experiments on the COIL-100 image library demonstrated the potential of the method by achieving 99.9% recognition rate for 18 training views per object. Even for a single training view, the correct model appear among the top four for almost 90% of the images. Robustness to severe occlusion was demonstrated by only a moderate decrease of recognition performance in an experiment where half of each test image was erased.

References

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

[2] Surrey object image library. http://www.ee.surrey.ac.uk/Research/VSSP/demos/colour/soil47.

[3] 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.

[4] F. Ennesser and G. Medioni. Finding waldo, or focus of attention using local color information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):805–809, August 1995.

[5] Y. Lambdan and H. Wolfson. Geometric hashing: A general and efficient model based recognition scheme. In *Proceedings of International Conference on Computer Vision*, pages 238–249, 1988.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Distinguished regions for wide-baseline stereo. Research Report CTU–CMP–2001–33, Center for Machine Perception, K333 FEE Czech Technical University, November 2001. ftp://cmp.felk.cvut.cz/pub/cmp/articles/matas/matas-tr-2001-33.ps.gz.

[7] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of International Conference on Computer Vision*, pages 525–531, 2001.

[8] J. L. Mundy and A. Zisserman, editors. *Geometric Invariance in Computer Vision*. The MIT Press, 1992.

[9] 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.

[10] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):872–877, May 1997.

[11] T. Suk and J. Flusser. Convex layers: A new tool for recognition of projectively deformed point sets. In F. Solina and A. Leonardis, editors, *Computer Analysis of Images and*

training views per object	18	8	4	2	1
recognition rate	92.6%	89.1%	82.6%	69.9%	63.3%

Table 4. COIL-100: Recognition rate for occluded images

Patterns : 8th International Conference CAIP'99, number 1689 in Lecture Notes in Computer Science, pages 454–461, Berlin, Germany, September 1999. Springer.

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

[13] T. Tuytelaars. *Local, Invariant Features for Registration and Recognition*. PhD thesis, University of Leuven, Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium, December 2000.

[14] T. Tuytelaars and L. V. Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *In Proceedings of British Machine Vision Conference*, pages 412–422, 2000.

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

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

A Affine Invariance of Covariance Matrix Construction

An affine transformation is a map $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form $F(\mathbf{x}) = A^T\mathbf{x} + \mathbf{t}$, for all $\mathbf{x} \in \mathbb{R}^n$, where A is a linear transformation of $\mathbb{R}^n$, assumed non-singular here. Let's consider a region Ω_1, and its transformed image $\Omega_2 = A\Omega_1$. Area of Ω_2 is given as

$$|\Omega_2| = \int_{\Omega_2} d\Omega_2 = \int_{\Omega_1} |A|\, d\Omega_1 = |A||\Omega_1|, \tag{2}$$

where $|A|$ is the determinant of A, and $|\Omega|$ is the area of region Ω. The center of gravity of region Ω is $\mu = \frac{1}{|\Omega|}\int_\Omega \mathbf{x} d\Omega$. The relation between the centers of gravity of transformed regions is:

$$\begin{aligned}\mu_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2} \mathbf{x_2}\, d\Omega_2 = \frac{1}{|A||\Omega_1|}\int_{\Omega_1} (A^T\mathbf{x_1} + \mathbf{t})|A|\, d\Omega_1 \\ &= A^T\frac{1}{|\Omega_1|}\int_{\Omega_1} \mathbf{x_1}\, d\Omega_1 + \frac{1}{|\Omega_1|}\int_{\Omega_1} \mathbf{t}\, d\Omega_1 \\ &= A^T\mu_1 + \mathbf{t} \end{aligned} \tag{3}$$

so the center of gravity changes covariantly with the affine transform. The covariance matrix Σ of a region Ω is a 2x2 matrix defined as $\Sigma = \frac{1}{|\Omega|}\int_\Omega (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\, d\Omega$. Covariance matrix of a transformed region Ω_2 is then

$$\begin{aligned}\Sigma_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2} (\mathbf{x_2}-\mu_2)(\mathbf{x_2}-\mu_2)^T\, d\Omega_2 \\ &= \frac{1}{|A||\Omega_1|}\int_{\Omega_1} (A^T\mathbf{x_1} + \mathbf{t} - (A^T\mu_1 + \mathbf{t})) \\ &\quad (A^T\mathbf{x_1} + \mathbf{t} - (A^T\mu_1 + \mathbf{t}))^T |A|\, d\Omega_1 \\ &= \frac{1}{|\Omega_1|}\int_{\Omega_1} (A^T(\mathbf{x_1}-\mu_1))(A^T(\mathbf{x_1}-\mu_1))^T\, d\Omega_1 \\ &= A^T\left(\frac{1}{|\Omega_1|}\int_{\Omega_1} (\mathbf{x_1}-\mu_1)(\mathbf{x_1}-\mu_1)^T\, d\Omega_1\right) A \\ &= A^T\Sigma_1 A \end{aligned} \tag{4}$$

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

$$\begin{aligned}\Sigma_2 &= U_2^TR_2^TR_2U_2 = \\ &\quad A^T\Sigma_1 A = A^TU_1^TR_1^TR_1U_1A \end{aligned} \tag{5}$$

thus

$$\begin{aligned}R_2U_2 &= R_1U_1A \\ U_2 &= R_2^{-1}R_1U_1A = RU_1A \end{aligned} \tag{6}$$

Hence the triangular matrix U, obtained through 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.