The Challenge of Image and Video Retrieval (CIVR2002)
July 18-19, 2002, London, UK

Local Affine Frames for Image Retrieval

Štěpán Obdržálek[1,2] and Jiří Matas[1,2]

[1] Center for Machine Perception, Czech Technical University, Prague, CZ
[2] Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK

Abstract. A novel approach to content-based image retrieval is presented. The method supports recognition of objects under a very wide range of viewing and illumination conditions and is robust to occlusion and background clutter. Starting from robustly detected 'distinguished regions' of data dependent shape, local affine frames are established by affine-invariant constructions exploiting invariant properties of the second moment matrix and bi-tangent points. Direct comparison of photometrically normalised colour intensities in normalised frames facilitates robust, affine and illumination invariant, but still very selective matching.
The potential of the proposed approach is experimentally verified on FOCUS — a publicly available image database — using a standard set of query images. The results obtained are superior to the state of the art. The method operates successfully on images with complex background, where the sought object covers only a fraction (around 2%) of the database image. Examples of precise localisation of the query objects in an image are shown too.

1 Introduction

In recent few years, the number of digital images that a user could search increased rapidly, fueling the need for content-based image retrieval systems. Many approaches addressing the problem of image retrieval were introduced, the most common being those using colour histograms [1–3], texture [4], shape [5–7], colour invariants [8, 9] or graph representations of colour content [10]. For a comprehensive survey, see [11].

In this paper we focus on a class of retrieval problems where the query depicts (a part of) an object of interest. We assume that the query object may cover only a franctional part of the database image and that it may be viewed from a significantly different viewpoint and under different illumination.

The proposed approach is based on robust, affine and illumination invariant detection of local affine frames (local coordinate systems). Local correspondences between the query and database images are established by a direct comparison of normalised colour in image patches with shape normalised according to the affine frames. The method achieves the discriminative power of template matching while maintaining the invariance to illumination and object pose changes of techniques using more general feature descriptors. In addition, for every local correspondence obtained, the local inter-image transformation is known, making it possible to robustly localise the query in the database image.

⋆ The authors were supported by the EU project IST-2001-32184, the Czech Ministry of Education project MSM 210000012 and a CTU grant No. CTU0209613.

The most closely related work is that of Tuytelaars and Gool [9], where local regions were also affine-invariantly found, but these regions were used to determine the image area over which affine moment invariants were computed. We argue here, that once image regions are found in a affine-invariant way, matches can be established by direct comparison of intensity profiles over these regions.

The main contribution of the paper is the utilisation of several affine-invariant constructions of local affine frames (LAFs) for determination of local image patches being put into correspondence. Robustness of the matching procedure is accomplished assigning multiple frames to each detected image region, and not requiring all of the frames to match. The outline of the proposed retrieval process is as follows:

1. For every database and query image compute distinguished regions, establish local affine frames, and generate intensity representation of local image patches normalised according to the local frames.
2. Establish correspondences between frames of query and database images, directly comparing the local image intensities. Estimate the match score based on the number and quality of established correspondences.
3. Combine affine transformations provided by every matched frame pair to establish an estimate of query location in the database image.

The paper is organised as follows. In Section 2 we briefly review the concept of distinguished regions. Section 3 gives a description of procedures giving local affine frames on distinguished regions of complex shapes. Section 4 details how correspondences between the local affine frames are established, and 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 [12], 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 [12] which includes a formal definition of the MSERs and a detailed description of the extraction algorithm. The report is available online. Examples of detected MSERs are shown in Figure 1. Note that DRs do not form segmentation, since 1. DRs do not cover entire image area, and 2. DRs can be (and usually are) nested.

Fig. 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 1. selecting only stable frames and 2. employing multiple processes for frame computation.

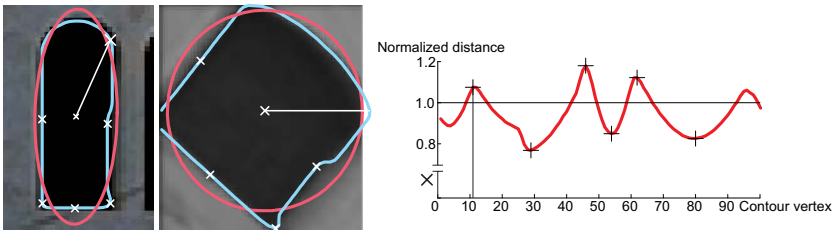

Fig. 2. Construction of affine frames. Original image, shape-normalised image and normalised distances between the center of gravity and contour

Frame constructions. Two main groups of affine-invariant constructions are proposed, one based on region normalisation by the covariance matrix (see Appendix A for definition and proof of affine invariance), second on 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 contour point of extremal (either minimal or maximal) distance from the CG 2. CG to 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 concavity of the region.

In frame constructions derived from the bi-tangants (the line segments of convex hull bridging concavities), the two tangent points are combined with a third point to

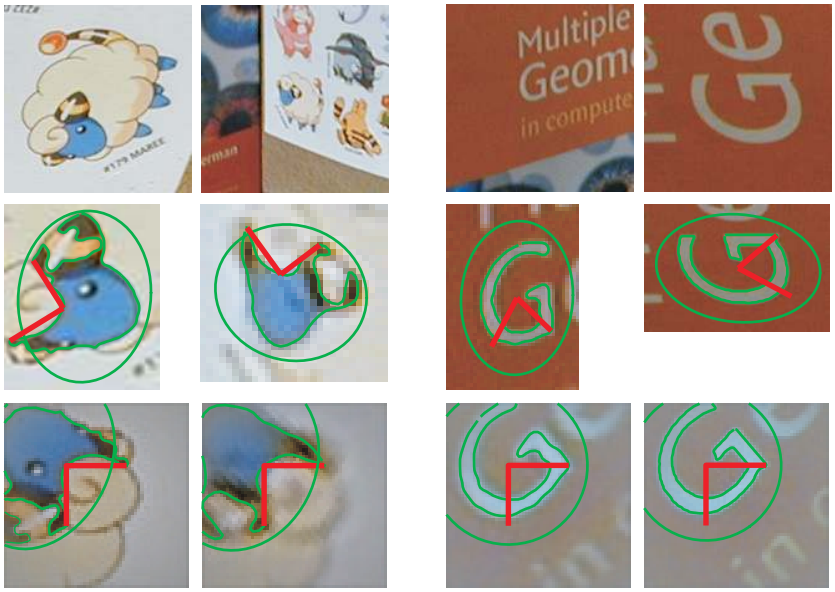

Fig. 3. Construction of affine frames. top row - original views, middle row - detected frames, bottom row - normalised frames

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 done 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 a part of the DR (containing the concavity).

Affine covariance of the center of mass and of the covariance matrix is shown in Appendix A. The invariance of the bi-tangets is a consequence of the affine invariance (and even projective invariance) of the convex hull construction [13, 14]. The invariance of the maximal-distance-from-a-line property is easily appreciated taking into account that affine transform maintains parallelism of lines and their ordering.

The process of extremal point detection is visualised in Figure 2. A region detected in an image (left) is transformed to the shape-normalised frame (middle). Normalised distances of contour points are plotted on the right. The ellipse defined by the covariance matrix of the region is transformed to the unit circle in the normalised frame. To complete affine frames, directions to the extremal points are used to resolve the unknown rotation.

Figure 3 shows an example of a construction of local affine frame for two objects. In the leftmost two columns, frames are build from bi-tangent points and the points most distant from the bi-tangents. The two rightmost columns show frame construction

based on DR's covariance matrix and extremal distance points. The top row displays two different views of the same object, detected frames are shown in the middle row, and normalised frames are depicted in the bottom row.

4 Matching

As a final step of our method, a matching score is computed between a pair of images. Since local affine frames have been established on distinguished regions, (geometrically) invariant descriptors of local appearance are not needed. The similarity assessment can rely simply on correlating photometrically normalised regions defined intrinsically in terms of local coordinate frames.

Figure 4 shows an example of normalised frames for a pair of images being put into correspondence, demonstrating some desirable properties of the LAF method. The object of interest undergoes a full affine change (anisotropic scale, rotation, skew), it is partially occluded, and covers only about 5% of the database image. However, many of the detected LAFs cover the same part of the object surface, which is clearly seen in the right part of Figure 4.

Having two sets of local affine frames, set S_1 of frames computed on the query image, and set S_2 on a database image, the computation of the matching score can be outlined as follows:

1. For every frame $f_{1i} \in S_1$ find such a frame $f_{2i} \in S_2$ so that f_{1i} and f_{2i} are of the same frame type, and that the intensity-normalised distance $d_i = |f_{1i}, f_{2i}|$, $d_i \in \langle 0, 1 \rangle$ is minimal, ie. $d_i = min|f_{1i}, f_2|, \forall f_2 \in S_2$
2. matching score $m = \sum_i (1 - d_i)^2$

Considering only the best match for every query frame makes the matching score independent of the number of frames defined on individual database images.

5 Experiments on the FOCUS database

We tested the retrieval performance of the proposed method on the FOCUS image database, containing 360 colour high-resolution images of commercials scanned from miscellaneous magazines. For comparison purposes, we run an experiment with an identical setup as the SEDL system introduced by Cohen [15]. 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:

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

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.

In Table 1, 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

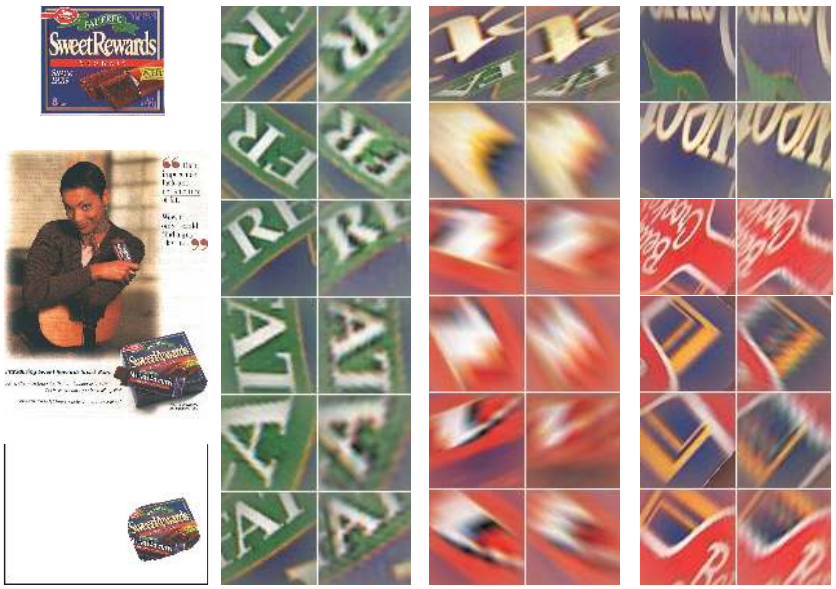

Fig. 4. Samples of correspondences established between frames of query (left columns) and database (right columns) images. On the left, the query, the database image, and the query localisation is shown.

Table 1. Retrieval performance compared to the SEDL system.

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%

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 (1). 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 proposed 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 5 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. Adapting the method for categorisation task is an open problem, possible approaches include adopting more flexible local representations, eg. local colour and texture distributions or flexible local templates.

Query localisation. Since the matching establishes explicit correspondences and mappings between parts of the query and the database image, it is possible to localise precisely the position of the query in the database image. The process is demonstrated in Figure 6. On the left, the query and database images are shown. Every individual cor-

Fig. 5. Examples of query (left) and database images (right) not retrieved in the FOCUS experiment

respondence of frames provides a single estimate of the query location in the database image. These estimates are displayed as white parallelograms in the central image. A voting scheme is applied to accumulate the estimates to form a clipping mask, shown in the fourth image. The clipped-out part of the database image, where the query was located, is shown on the right of Figure 6. Two other examples of query localisation are presented in Figure 7.

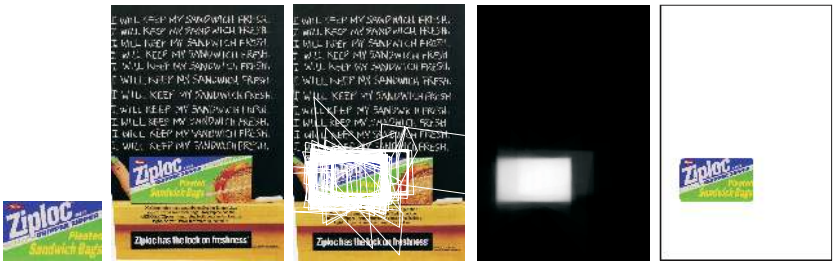

Fig. 6. Query localisation in the database image

Fig. 7. Sample query localisation results, query images, database images, and query localisations

6 Conclusions

In this paper, a novel procedure for image retrieval was introduced. Starting from robustly detected distinguished regions of data dependent shape, local affine frames were

obtained. The constructions of affine frames was proved affine covariant, and experimentally shown to be stable. Direct comparison of photometrically normalised colour intensities in normalised frames allowed for robust and selective matching. Fully determined frame to frame correspondences made it possible to robustly localise the occurrence of the query in the retrieved database image.

Experimental results obtained on a publicly available image database, a recall of 83% and a precision of 93% were superior to the state of the art [15].

References

1. Swain, M., Ballard, D.: Color indexing. In: International Journal of Computer Vision, vol. 7, no. 1. (1991) 11–32
2. Finlayson, G.D., Chatterjee, S.S., Funt, B.V.: Color angular indexing. In: ECCV. (1996) 16–27
3. Funt, B., Finlayson, G., Color, C.: Color constant color indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995) 522–529
4. Liu, F., Picard, R.W.: Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996) 7–733
5. E., G., R., M.: Shape similarity-based retrieval in image database systems. SPIE **1662** (1992) 2–8
6. Mokhtarian, F., Abbasi, S., Kittler, J.: Robust and efficient shape indexing through curvature scale space. In: In Proceedings of British Machine Vision Conference, Edinburgh, UK. (1996) 53–6
7. Bimbo, A., Pala, P.: Effective image retrieval using deformable templates. In: ICPR96. (1996) 120–123
8. Mindru, F., Moons, T., Gool, L.V.: Recognizing color patterns irrespective of viewpoint and illumination. In: CVPR99. (1999) 368–373
9. Tuytelaars, T., Gool, L.V.: Content-based image retrieval based on local affinely invariant regions. In: Proc. Visual '99: Information and Information Systems. (1999) 493–500
10. Park, K., Yun, I., Lee, S.U.: Color image retrieval using a hybrid graph representation. Journal of Image and Vision Computing **17(7)** (1999) 465–474
11. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
12. Matas, J., Chum, O., Urban, M., Pajdla, T.: Distinguished regions for wide-baseline stereo. Research Report CTU–CMP–2001–33, Center for Machine Perception, K333 FEE Czech Technical University, Prague, Czech Republic (2001) ftp://cmp.felk.cvut.cz/pub/cmp/articles/matas/matas-tr-2001-33.ps.gz.
13. Suk, T., Flusser, J.: Convex layers: A new tool for recognition of projectively deformed point sets. In Solina, F., Leonardis, A., eds.: Computer Analysis of Images and Patterns : 8th International Conference CAIP'99. Number 1689 in Lecture Notes in Computer Science, Berlin, Germany, Springer (1999) 454–461
14. Mundy, J.L., Zisserman, A., eds.: Geometric Invariance in Computer Vision. The MIT Press (1992)
15. Cohen, S.: Finding color and shape patterns in images. Technical Report STAN-CS-TR-99-1620, Stanford University (1999)

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 transformes 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

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

Cholesky decomposition of a symmetric and positive-definite matrix Σ is a factorization $\Sigma = U^TU$, 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

$$\Sigma_2 = U_2^TR_2^TR_2U_2 = A^T\Sigma_1 A = A^TU_1^TR_1^TR_1U_1A \tag{5}$$

thus

$$R_2U_2 = R_1U_1A \qquad\qquad U_2 = R_2^{-1}R_1U_1A = RU_1A \tag{6}$$

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.