Image Retrieval Using Local Compact DCT-based Representation

Štěpán Obdržálek[1] 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. An image retrieval system based on local affine frames is introduced. The system provides highly discriminative retrieval of rigid objects under a very wide range of viewing and illumination conditions, and is robust to occlusion and background clutter. Distinguished regions of data dependent shape are detected, and local affine frames (coordinate systems) are obtained. Photometrically and geometrically normalised image patches are extracted and used for matching.
Local correspondences are formed either by direct comparison of photometrically normalised colour intensities in the normalised patches, or by comparison of DCT (discrete cosine transform) coefficients of the patches. Experimental results are presented on a publicly available database of real outdoor images of buildings. We demonstrate the effect of the number of DCT coefficients that are used for the matching. Using the DCT, excellent results with a retrieval performance of 100% in rank 1 are achieved, and memory usage is reduced by a factor of 4.

1 Introduction

The widespread availability of digital images, and the increasing ease of their acquisition, distribution and storage, give rise to miscellaneous applications demanding reliable retrieval of images from digital databases. Many approaches addressing the problem of image retrieval were introduced, the most common being those using global descriptors of whole images, like colour histograms [1, 2], texture [3], shape [4], or colour invariants [5, 6]. For a comprehensive survey, see [7]. We leave aside the problems of connecting the user's query specification to the image representation (the problem known as the 'semantic gap'), and focus on the class of retrieval problems where the query is formed by an image of (a part of) the object of interest. We assume that the query object may cover only a fractional part of the database image and that it may be viewed from a significantly different viewpoint and under different illumination.

Variations in an object's appearance caused by viewpoint and environment changes are generally complex. Objects with intricate shapes change their overall look dramatically even for small differences in viewpoints. To simplify the situation, we assume that these variations, although complex in general, can be reasonably well approximated by simpler transformations at local scale. Geometric image deformations are locally approximated by 2D affine transformations, photometric changes by affine transformations of individual RGB channels.

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 measurements

⋆ The authors were supported by European Union under project IST-2001-32184, by Czech Ministry of Education under project LN00B096, and by CTU grant No. CTU0307013.

in local image patches with shape and colour normalised according to the affine frames. The method compares well with the state of the art. Object recognition and retrieval results on standard public image databases COIL-100 (mostly man-made 3D objects, no object occlusion nor background clutter) and FOCUS (planar logos, no occlusion but significant background clutter) are superior to any published results. The experiments are described in detail in [8, 9].

The presented image retrieval system is motivated by a real application: the localisation of an user in an outdoor environment. The system handles real outdoor images where the illumination varies due to weather changes, where objects are occluded, and where the background is cluttered. As the size of the database of known objects increases, the memory requirement of the object representation becomes important. We use the discrete cosine transform (DCT) to efficiently encode the local intensity information. The memory usage is thereby reduced by a factor 4, while the retrieval performance is maintained or even slightly improved.

The rest of the paper is organised as follows. In Section 2 we present an overview of the retrieval process and briefly discuss the concepts of distinguished regions and local affine frames. Section 3 details how images are represented by a set of local measurements, and in Section 4 experimental results are presented. Finally, Section 5 presents the conclusions.

2 Overview of the Retrieval Process

The outline of the proposed retrieval process is as follows (the first three steps are visualised in Fig. 1):

1. For every database and query image compute distinguished regions.
2. Construct local affine frames on the regions.
3. Generate intensity representations of local image patches normalised according to the local affine frames.
4. Generate discrete cosine transformation (DCT) representations of the normalised local patches.
5. Establish correspondences between frames of query and database images, by computing the euclidean distance between the local image intensities or their DCT coefficients, and by finding the nearest match.
6. An estimate of the match score is based on the number and quality of the established local correspondences.

In the rest of this Section we briefly introduce the concepts of the first two steps, the distinguished regions and the local affine frames. Remaining steps are discussed in the following sections.

Distinguished Regions (DRs) are image elements (subsets of image pixels), that posses some distinguishing 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 [10], the *Maximally Stable Extremal Regions* (MSERs). 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

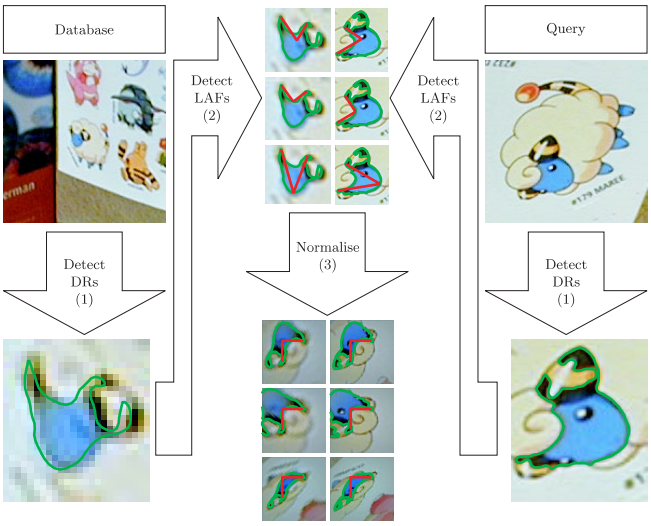


Fig. 1. Block diagram: obtaining the local, affine invariant image descriptors.

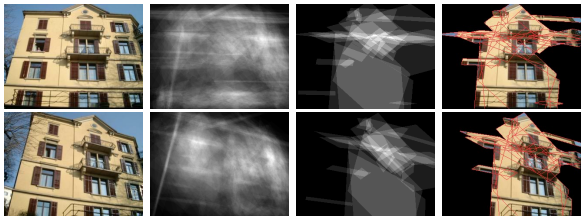
near real-time run time, and 4. since no smoothing is involved, both very fine and coarse image structures are detected. We do not describe MSERs here; the reader is referred to [10] which includes a formal definition of MSERs and a detailed description of the extraction algorithm.

Local affine frames (LAFs, local object-centered coordinate systems) allow normalisation of image patches into a canonical frame, and enable direct comparison of photometrically normalised intensity values, eliminating the need for invariants. For every distinguished region, multiple frames are computed. The actual number of the frames depends on the region's complexity. While simple elliptical regions have no stable frames detected, regions of complex non-convex shape may have tens of frames associated. Robustness of our approach is thus achieved by 1. selecting only stable frames and 2. employing multiple processes for frame computation. A detailed description of the local affine frame constructions is given in [9] and [8].

3 Image representation

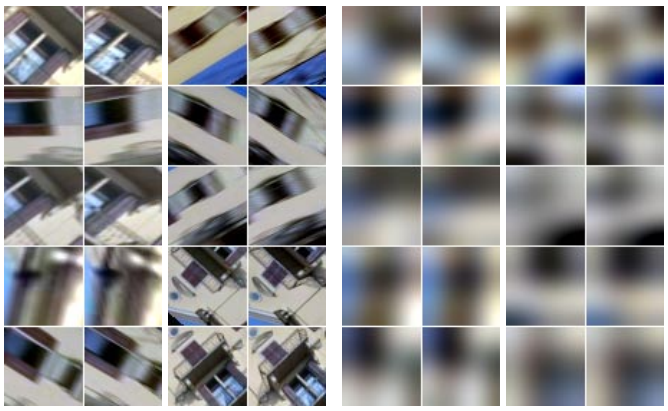
Images are represented by sets of local measurements. Since local affine frames are established, there is no need for geometrically invariant descriptors of local appearance. Any measurement taken relative to the frame is affine invariant.

Geometry. 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 an established frame F is described in homogenous coordinates by a 3 by 3 matrix $\mathbf{A}_F = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix}$. The image area (defined in terms of the affine frame) where the local measurements are taken from is referred to as a measurement region (MR). The choice of MR shape and size is arbitrary. Our choice is to use a square MR centered around a detected LAF,



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

Fig. 2. Coverage of images. (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 (including mismatches), (d) image area covered by the corresponding patches.



(a) (b)

Fig. 3. Examples of correspondences established between frames of query (left columns) and database (right columns), for the image pair from Figure 2. (a) geometrically and photometrically normalised image patches, (b) the same patches reconstructed from 10 DCT coefficients per colour channel.

specifically an image area spanning $\langle -2, 3\rangle \times \langle -2, 3\rangle$ in the frame coordinate system. See Figure 2 for an example of how an image is covered by measurement regions.

Photometry. Our model assumes a linear camera (ie. a camera without gamma-correction) and that no specular reflections are present in the local patches. The combined effect of different scene illumination and camera and digitiser settings (gain, shutter speed, aperture) can be then represented by affine transformations of individual colour channels. The transformation of intensities in colour channels between two corresponding patches I and I' is considered 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 constants $m_r, n_r, m_g, n_g, m_b, n_b$ differ for individual correspondences. This model agrees with the monochromatic reflectance model [11] and is an affine extension of the diagonal model, commonly used in colour constancy problems. To achieve invariance

to affine photometric variations and to enable direct intensity comparison, the patch intensities are transformed into a canonical form: the intensities in individual colour channels are affinely transformed to have zero mean and unit variance.

Normalisation Procedure. The normalisation of a local image patch proceeds in four steps:

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

The twelve normalisation parameters ($a_1 \dots a_6$, m_r, n_r, m_g, n_g, m_b, n_b) are stored along with the normalised intensity measurement. When considering a pair of patches for a correspondence, these twelve parameters are combined to provide the local transformation (both geometric and photometric) between the images. Constraints can be put here on the transformation to prune potential matches. Typical constraints may include: allowing only small scale changes for images taken from approximately constant distance from the objects, rejecting significant rotations when upright camera and object orientations can be assumed, allowing for only small illumination changes for images taken in a controlled environment, and many others. If the runtime conditions are known, the unconstrained invariance can so be traded for higher discriminativity.

Intensity representation. After the normalisation, any measurement on a local patch can be directly compared to another. No technique is necessary for further alignment of the potentially corresponding pairs (e.g. the maximisation of correlation over an unknown rotation). To establish correspondences between patches, we can use directly the underlying intensity function. The normalised MR content is stored in a discretised form as an array of 15×15 pixels, and the correlation coefficient is used as the similarity measure of two patches. See Figure 3a for an example of pairs of normalised patches.

Discrete Cosine Transformation Patch description by a discretised intensity function is high-dimensional. In many pattern recognition problems (eg. in face recognition) the Karhunen-Loeve (KL) transformation is used to reduce the feature dimensionality without significant deterioration of the recognition performance. The KL transformation has drawbacks though. Mainly, it depends on the second-order statistics of the training data, ie. the training (database) images have to be known in advance.

For the dimensionality reduction we therefore use the discrete cosine transformation (DCT) instead of the Karhunen Loeve transformation. The DCT has the following desirable properties:

- For uniformly distributed data, the DCT approximates the Karhunen-Loeve transformation [12].
- Fast algorithms exist that computes DCT with $O(n \log n)$ time complexity.
- Due to the widespread use of DCT in image and video compression domain (JPEG, MPEG, etc.), hardware implementations of DCT are widely available.
- Unlike the KL transformation, DCT does not require a training set.

Fig. 4. Example of database images. Five images are present for every of the 201 buildings.

Keeping only low frequency DCT coefficients, the high frequencies are neglected. Local patches differing only in the high frequencies become indistinguishable. On the other hand, the high frequencies are corrupted by image noise and by small misalignments caused by inexact frame detection. The DCT is thus less sensitive to the imprecisions present in the normalised patches, as is experimentally verified in Section 4. The number of DCT coefficients that should be used depends on the discriminativity required, ie. basically on the database size. In Section 4, we experimentally show how the number of the DCT coefficients affect the retrieval performance. In Figure 3b 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 3a.

4 Experiments

Dataset. The experiments were conducted on a set of images of 201 different buildings in Zurich, Switzerland. The dataset was kindly provided by ETH Zurich and is publicly available [13]. The database consists of five photographs of every of the 201 buildings. The photographs are taken from different viewpoints but under approximately constant illumination conditions. The database contains 1005 images in total, the image resolution is 320×240 pixels. Examples of the database images are shown in Figure 4. 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 5.

Experimental Protocol. 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, $N = \min(R, 5)$.

Fig. 5. Examples of corresponding query (left columns) and database (right columns) images. The image pairs exhibit occlusion, varying illumination and viewpoint and orientation changes.

Results. The two local patch representations (see Sect. 3) are compared, ie. the directly stored intensities versus the DCT coefficients. The results are summarised in Table 1. 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 line in Table 1 shows the results published in [14].

Summary. Generally, the proposed retrieval system performed well, obtained results were superior to results published in [14]. The retrieval performance was, or was close to, 100% in the first rank. The DCT representation performed slightly better than the direct intensity representation. We believe that this is due to the DCT properties discussed in Section 3 – 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.

5 Conclusions

In this paper, an image retrieval system based on local affine frames (object-centered coordinate systems) was presented. The system is robust to object occlusion and background clutter, and allows retrieval of objects in images taken from significantly different viewpoints. Normalised image patches are extracted, and photometrically and geometrically normalised according to the detected frames. Local matches are formed both by direct comparison of photometrically normalised colour intensities in the normalised patches, and by comparison of DCT (discrete cosine transform) coefficients of the patches. Both representations allow for robust and selective matching, providing excellent retrieval performance. Experimental results obtained on a publicly available im-

Method	Average recall r_R					Memory usage
	r_1	r_2	r_3	r_4	r_5	
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 [14]	86.1%					

Table 1. Summary of experimental results.

age dataset of buildings were superior to other published results. Retrieval performance of 100% in rank one was achieved when the local image patches were represented by 15 DCT coefficients in every colour channel. The DCT representation performed better in terms of recall rate and required about 5 times less memory storage than representation by the intensities of the normalised patches.

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. Liu, F., Picard, R.W.: Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. IEEE PAMI **18** (1996) 7–733
4. 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
5. Mindru, F., Moons, T., Gool, L.V.: Recognizing color patterns irrespective of viewpoint and illumination. In: CVPR99. (1999) 368–373
6. 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
7. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE PAMI **22** (2000) 1349–1380
8. Obdržálek, Š., Matas, J.: Object recognition using local affine frames on distinguished regions. In: The British Machine Vision Conference (BMVC02). (2002)
9. Obdržálek, Š., Matas, J.: Local affine frames for image retrieval. In: The Challenge of Image and Video Retrieval (CIVR2002). (2002)
10. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In Rosin, P.L., Marshall, D., eds.: Proceedings of the British Machine Vision Conference. Volume 1., London, UK, BMVA (2002) 384–393
11. Healey, G.: Using color for geometry-insensitive segmentation. Journal of the Optical Society of America **6** (1989) 86–103
12. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632 (1986)
13. Shao, H., Svoboda, T., Van Gool, L.: ZuBuD — Zurich Buildings Database for Image Based Recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology (2003)
14. Shao, H., Svoboda, T., Tuytelaars, T., Van Gool, L.: Hpat indexing for fast object/scene recognition based on local appearance. In: International Conference on Image and Video Retrieval. (2003) To appear.