Object Recognition using Local Affine Frames on Distinguished Regions

Štěpán Obdržálek[1,2] and Jiří Matas[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

A novel approach to appearance based object recognition is introduced. The proposed method, based on matching of local image features, reliably recognises objects under very different viewing conditions.

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.

The potential of the approach is experimentally verified on COIL-100 and SOIL-47 – publicly available image databases. On SOIL-47, 100% recognition rate is achieved for single training view per object. On COIL-100, 99.9% recognition rate is obtained for 18 training views per object. Robustness to severe occlusions is demonstrated by only a moderate decrease of recognition performance in an experiment where half of each test image is erased.

1 Introduction

Object recognition is one of the oldest fields in computer vision, and is still attracting the attention of many researchers. As a consequence, a wide range of approaches have been proposed. 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, without attempting high-level image analysis. 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,

*The authors were supported by the European Union project IST-2001-32184, the Czech Ministry of Education project MSM 212300013, The Grant Agency of the Czech Republic project GACR 102/02/1539 and a CTU grant No. CTU0209613.

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. Common approaches include geometric hashing exploiting geometric configuration of local image features [6], PCA analysis over local image areas (eigenwindows) [10], matching local colour histograms [13, 4], matching gaussian derivatives in neighbourhoods of scale-invariant interest points [11, 8], or matching moment invariants in local affine-invariant regions [15, 14].

In this work, an assumption is made that image deformations can be reasonably well approximated by local affine transformations of both the geometry and the 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 and illumination invariant detection of local affine frames (local coordinate systems). Local correspondences between individual images are established by a direct comparison of normalised colours in image patches represented canonically in normalised affine frames. The method achieves the discriminative power of template matching while maintaining the invariance to illumination and object pose changes.

The most closely related work is that of Tuytelaars [14], where local regions were also affine-invariantly found, but these regions were used to determine the image area over which moment invariants were computed. We argue here, that once image regions are found in an affine-invariant way, matches can be established by direct comparison of intensity profiles over these regions. Tuytelaars also proposed to establish correspondences using normalised correlation over the shape-normalised regions, but since the regions were determined up to an unknown rotation, a computationally expensive maximisation of the correlation was used.

The main contribution of the paper is the utilisation of several affine-invariant constructions of local affine frames (LAFs) for the determination of local image patches that are being put into correspondence. The robustness of the matching procedure is accomplished by assigning multiple frames to each detected image region, and not requiring all of the frames to match. Matching score is estimated as the number of established local correspondences, without enforcing a global model consistency. Ignoring the consistency generalises the object representation to such views where the global appearance substantially differs from the training views, but still some of the local features are preserved so that local correspondences can be established. However not rejecting the inconsistent matches requires that their fraction is low, putting thus high demands on the selectivity of the correspondence generation process.

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 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 [7], 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 [7] which includes a formal definition of the MSERs and a detailed description of the extraction algorithm. The report [7] 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.

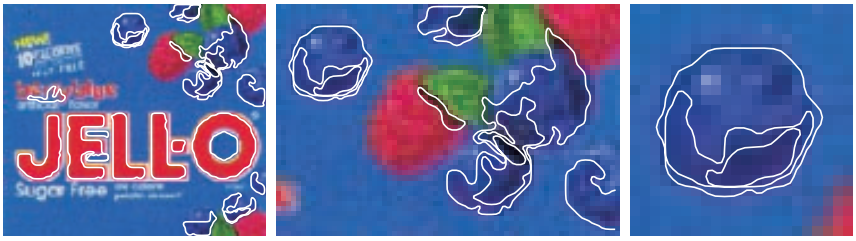


Figure 1: An example of detected distinguished regions of MSER type

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 $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 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 $\mu = \frac{1}{|\Omega|}\int_\Omega \mathbf{x} d\Omega$.

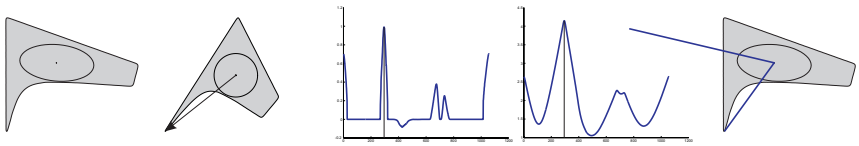
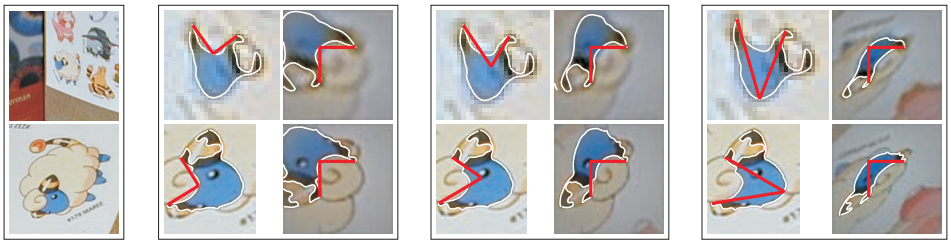


Figure 2: Construction of affine frames. From left to right: a distinguished region (the grey area), the DR shape-normalised according to the covariance matrix, normalised contour curvatures, normalised contour distances to the center of DR, and one of the constructed frames represented by its basis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

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 [16]. The invariance of the bi-tangents is a consequence of the affine invariance (and even projective invariance) of the convex hull construction [12, 9]. 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. A measurement region MR is defined in local coordinate systems of the affine frames, but the choice about MR 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 $\hat{I}(x,y) = (I(x,y) - \mu)/\sigma, \quad x,y \in \{1..21\}$ 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 the between-frame transformation (both geometric and photometric). An application-specific constraints can be applied here to prune the 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 controlled environment, and many others.

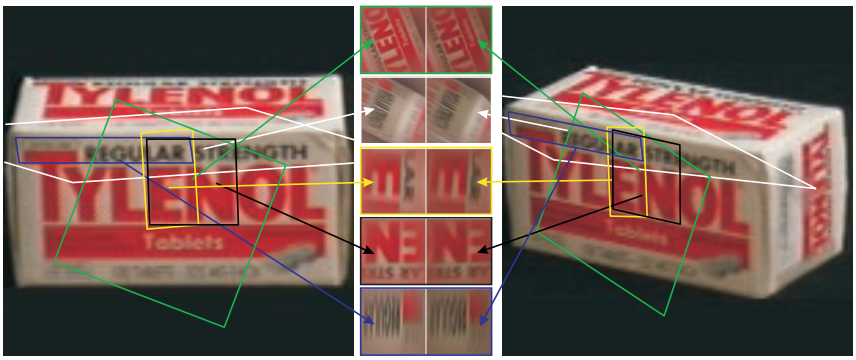


Figure 4: Examples of correspondences established between frames of a training image (left) and a test image (right).

Local correspondences are established by correlating $\hat{I}$s that invariantly represents colour measurements from the respective MRs. Figure 4 shows an example of correspondences found for a pair of images from the COIL-100 database. The choice of the best strategy for the computation of the inter-image 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 images 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) considered only as an affine 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.

5 Experiments on COIL-100 and SOIL-47 databases

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 5 several objects from the database are shown.

Figure 5: Several objects from COIL-100 database

Table 1 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 insensivity to image deformations. The LAF approach performs best in all experiments, regardless of the number of training views. For only four training views, the recognition rate is almost 95%, demonstrating the remarkable robustness to local affine distortions. In the case of 18 training views per object, only 5 out of the total 5400 test images were misclassified. Table 2 summarises achieved recognition rates up to rank 4. 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.

Average recognition time for a single image is 0.8 sec (on 1.4 GHz PC) for the case of four training views per object (i.e. matching every test image to 400 models). In order to evaluate the potential of the method, we have not implemented any kind of hypotheses pruning or indexing into our matching algorithm, evaluating all the correlations.

training views per object	18	8	4	2	1
total test views	5400	6400	6800	7000	7100
LAF training views	$0^\circ + k20^\circ$	$0^\circ + k45^\circ$	$45^\circ + k90^\circ$	$0^\circ, 90^\circ$	0°
LAFs	99.9%	99.4%	94.7%	87.8%	76.0%
SNoW / edges [17]	94.1%	89.2%	88.3%	-	-
SNoW / intensity [17]	92.3%	85.1%	81.5%	-	-
Linear SVM [17]	91.3%	84.8%	78.5%	-	-
Spin-Glass MRF [3]	96.8%	88.2%	69.4%	57.6%	49.9%
Nearest Neighbour [17]	87.5%	79.5%	74.6%	-	-

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

Due to the symmetric nature of many objects in the COIL-100 database, not only the number of training views, but also the selection of viewing angles for the training affects

Rank	# of training views per object				
	18	8	4	2	1
$= 1$	99.9%	99.4%	94.7%	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 2: COIL-100: Recognition rate, ranks 1 to 4

the recognition rate. An example is illustrated in figure 7 for the case of two training views. The pictures show the distribution of recognition rate over test view angles, accumulated for all the objects. In the left image, single training view at 0° was used – the recognition rate is 76%. The increased recognition rate in the 180° views is an indication that many objects appear similar when viewed from the opposite side. Adding a second training view at 180° (middle image), the recognition rate increases to 81.4%, and adding a training view at 90° (right image), the recognition rate reaches 87.8%. In the case of a single training view, the recognition rate varies from the worst 68.4% (for 270° training) to the best 77.3% (for 155°), with the average value being 74.1%.[1] The potentially interesting issue of selecting optimal set of views for each object was not pursued.

Figure 6: Several object from the SOIL-47 database

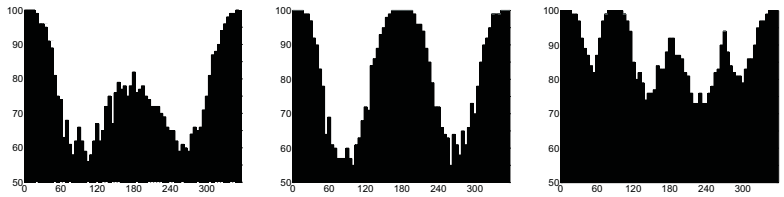
SOIL-47. We have also performed experiments on another publicly available image database, SOIL-47 [2]. Figure 6 shows few objects from the database. We have used identical experiment setup as in [5], i.e. using the same subset of 24 box-like objects, using one training view per object, and test view angles differing up to 45° (in [5] referred as views 6–15). The training images had the resolution twice as high as the test views. Our method achieved 100% recognition. Results are summarised in Table 3.

Method	LAFs	Graph matching	Geom. Hashing	Geom. Alignment
Recognition rate	100%	73%	63%	50 %

Table 3: SOIL-47: Recognition rate, in comparison to other methods [5]

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

[1] COIL-100 experiments are therefore repeatable (and comparable) only if training views are listed; the number of training views is insufficient



(a) (b) (c)

Figure 7: The effect of choice of training views: recognition rate as a function of the test view angle for (a) a single training view at 0°, (b) two training views at 0° and 180°, and (c) two training views at 0° and 90°

different numbers of training views. Figure 8 shows examples of the occluded test images. Recognition rates are summarised in Table 4.

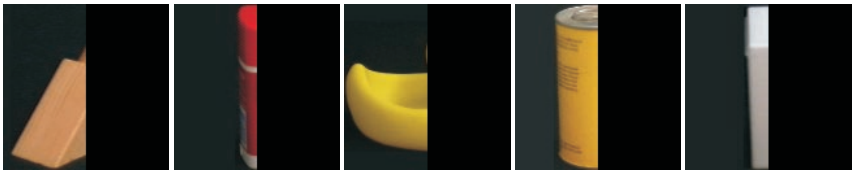


Figure 8: COIL-100: Examples of test images for the occlusion experiment

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

6 Conclusions

In this paper, a novel procedure for appearance based object recognition was introduced. 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 inter-image correspondences. Selective matching at the level of local features enabled successful recognition of object even when the objects were seen from angles differing by 180° from the training view. 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 occlusions was demonstrated by only a moderate decrease of recognition performance in an experiment where half of each test image was erased. In experiments on the SOIL-47 database, 100% recognition rate was achieved when using single training view and test views differing up to 45°.

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] Josef Kittler and Alizera Ahmadyfard. On matching algorithms for the recognition of objects in cluttered background. In *4th International Workshop on Visual Form IWVF2001*, pages 51–66, 2001.

[6] 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.

[7] Jiří Matas, Ondřej Chum, Martin Urban, and Tomáš 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.

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

[9] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision*. The MIT Press, 1992.

[10] 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.

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

[12] Tomas Suk and Jan Flusser. Convex layers: A new tool for recognition of projectively deformed point sets. In Franc Solina and Aleš Leonardis, editors, *Computer Analysis of Images and Patterns : 8th International Conference CAIP'99*, number 1689 in Lecture Notes in Computer Science, pages 454–461, Berlin, Germany, September 1999. Springer.

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

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

[15] Tinne Tuytelaars and Luc Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *In Proceedings of British Machine Vision Conference*, pages 412–422, 2000.

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

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