Preprint. Presented at ICPR 2006 in Hong-Kong.

FAIR: Towards A New Feature for Affinely-Invariant Recognition

Radim Šára, Martin Matoušek
Czech Technical University
Center for Machine Perception
Karlovo nam 13, CZ-12135 Prague, Czech Republic
{sara,xmatousm}@cmp.felk.cvut.cz

Abstract *In this paper we propose the first version of FAIR, a low-dimensional image neighborhood descriptor that shows performance comparable to SIFT introduced by Lowe. The dimension of FAIR we tested is 30, compared to the dimension of 128 in SIFT. Sensitivity of the FAIR descriptor to skew, rotation, image blur and noise is similar to SIFT. FAIR shows better localization in scale-space than SIFT. Several extensions of FAIR that could improve its performance are discussed.*

1 Introduction

Local image representations are an important component for establishing reliable and robust image correspondences in wide-baseline matching, panoramic image stitching, image retrieval and video mining, recognition, robot localization and obstacle avoidance, range image registration, etc. Basic types of complex neighborhood representations include SIFT [5] and its variant PCA-SIFT [2], Spin Images [3], GLOH [7], Shape Context [1], MSER [6] and a number of others. A comprehensive review and a comparison of state-of-the art point descriptors based on these representations is given in [7].

This paper proposes an efficient descriptor that is affinely quasi-invariant, shows good performance and has a relatively low dimension (30). Around identity, its sensitivity to affine image domain transformation and to image blur is similar to that of SIFT but its sensitivity to noise is better than in SIFT that was shown to be among the best descriptors for textured scenes [7]. FAIR is sector-based but has more sectors than SIFT which implies FAIR has the potential to improve the estimate of the neighborhood orientation (rotation) during matching.

The next section describes the construction of FAIR, Sec. 3 presents a thorough sensitivity testing and comparison to SIFT and Sec. 4 concludes the paper by discussing possible improvements.

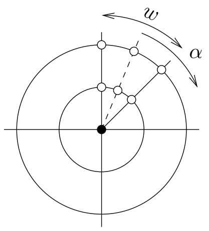

Fig. 1. The neighborhood representation of an interest point (black) consists of a star-like configuration of points over a set of concentric shells (left). The neighborhood is formed by a sector of fixed width w rotated by angle α in fixed-size increments giving a curvilinear descriptor.

2 The Elements of FAIR

The FAIR descriptor includes three components: (1) an image point neighborhood representation, (2) an invariant measurement and (3) a metafeature.

The neighborhood is somewhat similar to both spin image [3] and SIFT [5] and is represented by an angular sector of width w that is rotated around the central pixel in angular increments $d\alpha$, see Fig. 1. In practice, the neighborhood can be adapted by corrective affine mapping that is obtained from affine adaptation [4].

The next component is the affinely invariant measurement. Let $\mathbf{x}$ be a point in image domain and let $\nabla f(\mathbf{x})$ be image gradient at that point. Under affine mapping $\mathbf{A}$ the point transforms $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, so the gradient transforms $\nabla f(\mathbf{x}) \mapsto \mathbf{A}^{-\top}\nabla f(\mathbf{x})$. As shown in the appendix, the product

$$m_1(\mathbf{x}) = \mathbf{x}^\top \nabla f(\mathbf{x}) \qquad (1)$$

does not change under any non-singular affine mapping. Hence, we propose the invariant measurement to be $m_1(\mathbf{x})$. Invariance to brightness change b in $f(\cdot) \mapsto af(\cdot) + b$ is given by the fact we use image gradient, invariance to contrast a in will be obtained through normalization described below.

The last component of FAIR is a metafeature defined over the elementary affinely invariant measurements. Let the neighborhood sectors be indexed by their bisector angle α. As in SIFT, one can collect a histogram of elementary

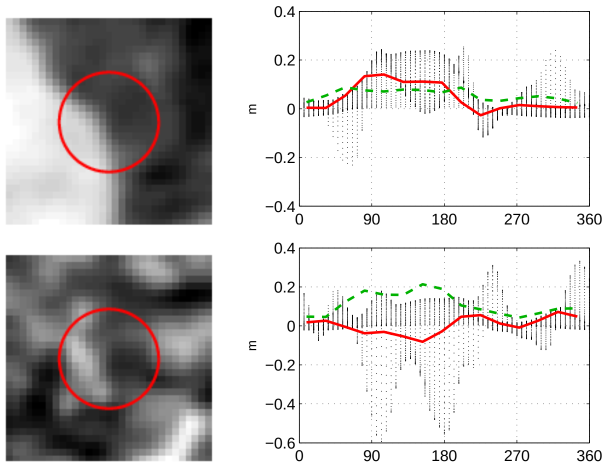

Fig. 2. Typical FAIR descriptor at two image points. The red line is $\mu(\alpha)$, the dashed dark green line is $\sigma(\alpha)$ and the points are the raw measurements $m_1(\mathbf{x})$ plotted as a function of α. The FAIR neigborhood has radius of 15 pixels (red circle).

measurements over the sector to be that sector's descriptor. But since the sector is smaller than in SIFT, it does not contain a sufficient number of independent measurements, we therefore represent it by a collection of sample statistics: In this paper we use the sample mean and variance. Hence, we compute the mean $\mu(\alpha)$ and the standard deviation $\sigma(\alpha)$ of all elementary measurements $m_1(\mathbf{x})$ over all points $\mathbf{x}$ that fall in the respective sector.[1] We will call the pair of curves $(\mu(\alpha),\, \sigma(\alpha))$ a descriptor of an image point. Fig. 2 shows descriptors of two different points in a real image.

To gain an insight into affine invariance of FAIR, suppose for a moment that the neighborhood consisted from only a single 360° sector and that we used a full histogram of that sector. In addition, suppose the neighborhood was infinite and the sampling was continuous (infinitely dense). In such case the entire histogram would be affinely invariant. The invariance has two sources. First, the individual measurements are affinely invariant and second, under affine mapping, which is bijective, a point in the neighborhood maps to another point in the neighborhood, i.e. no point escapes nor enters the neighborhood and we collect the same histogram. Once the sectors are smaller than 360°, some of the points can escape from their original sector and enter another. In such case theoretical invariance is lost. But the loss can be compensated during similarity computation. This is done by relative rotation of the two descriptors by an angle. Under such compensation we re-establish rotational and scaling[2] invariance. Invariance to skew could be re-established by warping the signals $(\mu(\alpha),\, \sigma(\alpha))$ onto each other, see the end of Sec. 2 for further discussion.

The version of FAIR that uses measurement $m_1(\mathbf{x})$ given by (1) will be called FAIR-1. The collection of image values $m_0(\mathbf{x}) = f(\mathbf{x})$ can also be used to construct a FAIR descriptor, the construction being exactly the same as FAIR-1. We will call this version FAIR-0. The FAIR-0 has the same sources of invariance as discussed in the previous paragraph. As shown in Sec.3, FAIR-0 is inferior to FAIR-1.

The invariance of the measurement m_1, m_0, together with the splitting of the neighborhood to large sectors are responsible for small sensitivity of FAIR to affine transformations of discrete image domain. The complexity of descriptor curves is responsible for the discriminability[3] of FAIR.

Note that due to sectoring, rotational invariance of the descriptor is lost, as in SIFT. One either has to use a alignment procedure as in [5] or design a similarity measure that includes the alignment.

Given an image, the complete procedure for computing FAIR features for matching proceeds as follows:

1. Select interest points $\mathbf{x}_j$, $j = 1, 2, \dots m$ (IP) and their natural scale by e.g. the DoG detector as in [5].
2. Find affinely-invariant neighborhood $\mathbf{S}_j$ for each IP as in [4]. The affine correction $\mathbf{S}_j^{\frac{1}{2}}\mathbf{x}$ is used to transform the FAIR neighborhood shown in Fig. 1.
3. Determine the natural orientation of the neighborhood, e.g. based on image gradient distribution, as in SIFT [5].
4. For each IP $\mathbf{x}_j$ at its natural scale and orientation: Collect the mean value $\mu(\alpha_i)$ and standard deviation $\sigma(\alpha_i)$ from measurements $m(\mathbf{x}_j)$ over each sector α_i. This gives vectors $\mu(\mathbf{x}_j) = [\mu(\alpha_i)]_{i=1}^{s}$ and $\sigma(\mathbf{x}_j) = [\sigma(\alpha_i)]_{i=1}^{s}$ where s is the number of sectors. Record the descriptor $\mathbf{D}(\mathbf{x}_i)$ which is a pair of normalized vectors (a 2 × s matrix)

$$\mathbf{D}(\mathbf{x}_i) = \left(\frac{\mu(\mathbf{x}_i)}{\|\mu(\mathbf{x}_i)\|},\, \frac{\sigma(\mathbf{x}_i)}{\|\sigma(\mathbf{x}_i)\|} \right). \tag{2}$$

To construct the FAIR descriptor, we use $s = 30$ sectors, each 24° wide, i.e. the angular increment is $d\alpha = 12°$. The overlap of the sectors acts as a filter on the descriptor and helps improve invariance breaking due to measurements exiting/entering a sector. The FAIR descriptor thus consists of 60 scalar values. Note that SIFT is represented by 128 scalar values. The diameter in pixels of the FAIR neighborhood is 15 pixels, which corresponds to 16 × 16 pixel neighborhood used by SIFT.

[1] Other statistics like the maximum and the minimum can be used but they are more sensitive, especially to image blur.

[2] As long as the neighborhood remains infinite.

[3] In newer literature discriminability is sometimes called 'distinctiveness.'

The reason we normalize in (2) is two-fold: We not only achieve invariance to image contrast but, as we have observed, we *significantly* improve insensitivity to image blur. This can be explained by the fact that the magnitude of image gradients decreases with blurring and that the normalization compensates the loss. This observation also strengthens the case for normalization in SIFT [5].

The feature distance between interest points $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d(\mathbf{x},\, \mathbf{y}) = \frac{1}{\sqrt{2}} \|\mathbf{D}(\mathbf{x}) - \mathbf{D}(\mathbf{y})\|_F \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. The distance is a harmonic mean of the vector norms for the individual components of the descriptor and it falls in the interval $0 \leq d(\mathbf{x},\, \mathbf{y}) \leq 2$. This is consistent with the distance recommended for SIFT [5].

Note that affine transformation does not change the order given by α. This means that under affine transformation the FAIR curves $\mu(\alpha)$, $\sigma(\alpha)$ warp by a monotonic function.[4] Future work includes a dynamic programming based warping of FAIR curves that could result in descriptors that work under non-rigid transformations as well.

3 Experiments

We compare both FAIR-1 and FAIR-0 to SIFT. Our goal is to compare the performance of these descriptors itself, not of the whole IP location, orientation normalization, similarity computation and matching pipeline. We therefore use DoG interest points at their natural scale but we neither detect the affinely invariant image neighborhoods, nor normalize the image neighborhood orientations nor warp the FAIR descriptors, i.e. we omit Steps 2 and 3 in the above procedure. The fixed-size image neighborhoods of IPs are thus considered independent samples from the 'world of all images.' This way we are able to observe design parameters of the descriptor itself.

In each test case there are two images: the original image I and its synthesized version $\tilde{I}$ altered by the corresponding transformation. Interest points are located in I and mapped by known affine mapping $\mathbf{A}$ to a sub-pixel location in the other image.

Distance $d(\mathbf{x}_i, \tilde{\mathbf{x}}_i)$ is evaluated for all IPs $i = 1, 2, \ldots, n$, where $\mathbf{x}_i$ is in the domain of I and $\tilde{\mathbf{x}}_i = \mathbf{A}\mathbf{x}_i$ is in the domain of $\tilde{I}$. From all pairs $(\mathbf{x}_i,\, \tilde{\mathbf{x}}_i)$ we construct an $n \times n$ distance matrix $\mathbf{C}$. In this matrix we count the number t of cases when both the row and the column minimum falls onto the diagonal. In a good descriptor, all n cases falls on the diagonal and $t = n$. In a descriptor that has poor performance, some of the cases fall off the diagonal. We therefore measure an 'overall performance ratio'

$$r = \frac{t}{n} \tag{4}$$

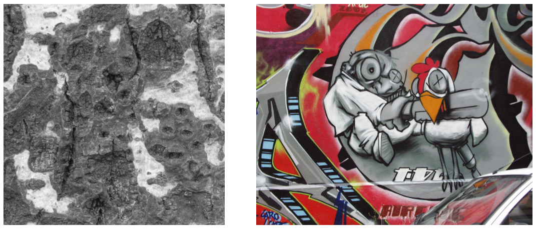
(a) textured (b) structured

Fig. 3. Test data: textured (a) and structured images (b) [7].

of the descriptor as the ability to recognize correspondences by nearest-neighbor search. A good descriptor has $r = 1$. This ratio is influenced by both discriminability of the descriptor and insensitivity to degradation factors.

To capture finer differences and decouple invariance from discriminability, we proceed as follows. From the n values on the diagonal of $\mathbf{C}$ we compute the mean e_D and standard deviation s_D. Then we set all diagonal elements of $\mathbf{C}$ to infinity. A set of all extremal pairs $E = \{(i, j)\}$ such that $\mathbf{C}(i, j)$ has the lowest value over the corresponding row i and column j is collected. Finally, we compute the mean e_E and standard deviation s_E over E.

The e_E estimates the distance to the nearest neighbor in feature space containing all possible measurements. This space has spherical topology due to normalization of the descriptors to unit vectors. A descriptor has a good *discriminability* if the e_E is large. The s_E estimates the uniformity of the density of all measurements in the feature space. Hence, a good-discriminability descriptor also has good s_E.

A descriptor has good *invariance*[5] if e_D and e_E are wide apart, significantly wider than their standard deviations s_D, s_E. The standard deviation s_D is small when the affine mapping does not redistribute the points in feature space non-uniformly. Hence, s_D should be small in a descriptor of good invariance.

Note that at the point when $e_E = s_E$ the expected performance r is about 50%. This is corroborated by the plots shown in this section.

Data used in the following experiment are shown in Fig. 3. Following the methodology of [7], we selected the two images to cover the character of typical scenes. The images are converted to gray-scale and normalized to the interval $[0, 1]$. We have used $n = 1500$ IPs per image.

Performance is evaluated on each image independently and the results are averaged over the two images. The tested descriptor does not know the actual affine mapping, blur or noise level.

We study four variants of image descriptors: FAIR-0 and FAIR-1 as defined above, SIFT, which is the standard SIFT

[4] The function is in fact even more constrained under affine mapping.

[5] We use the term 'invariance' loosely, as a shortcut for 'low sensitivity.'

feature with the rotational normalization of the neighborhood switched off and SIFT-O in which we left the rotational normalization on, as in the standard implementation of SIFT.

The result of each experiment is shown in five plots (see, e.g. Fig 4): The top wide plot shows the overall performance ratio r as a function of a degradation factor under study, the bottom four small plots are discriminability e_E (red curve) and sensitivity to the distortion e_D (blue curve), together with their respective standard deviations s_E and s_D (dashed). The ranges of e_E, e_R are $[0, 2]$ for both FAIR and SIFT, see (3), and the values are comparable since essentially the same norm is used in both SIFT and FAIR. The red curve is higher if the descriptor has greater discriminability. The blue curve is higher for any given factor's strength if the descriptor has greater sensitivity under the factor. Note, however that it is not possible to directly compare discriminabilities and sensitivities of two descriptors of different dimension. Hence, comparison of discriminability e_E and invariance e_D is not possible between SIFT and FAIR in this experiment. We can only compare FAIR-0 with FAIR-1 and SIFT with SIFT-O.

In all plots in Fig 4–6 we see that the discriminability of FAIR-1 is better than that of FAIR-0 (the red curve is higher in FAIR-1) and that discriminability of SIFT-O is slightly worse that in SIFT. Discriminability results are (and should be) almost unaffected by the type of degradation.

Fig. 4 compares the influence on the performance of rotation $\mathbf{A} = \mathbf{R}(\phi)$ and small shift of the interest point. Fig. 5 show results under scale change $\mathbf{A} = s\mathbf{E}$ and under Gaussian image blur of standard deviation σ. Except for the rotation, the remaining three tests are related to localization in scale-space. In the top-right plot of Fig. 4 and the top-left plot in Fig. 5 we see FAIR localizes with greater precision than SIFT.

The plots in Fig. 6 compare the influence of skew in the form

$$\mathbf{A} = \begin{bmatrix} 1 & q \\ 0 & 1 \end{bmatrix}$$

and Gaussian noise with standard deviation σ. The noise is applied to the input image. In skew, we again see similar performance of FAIR-1 and SIFT.

In noise, FAIR-1 shows performance comparable to SIFT. Although FAIR-0 and FAIR-1 have similar sensitivity to noise (blue curves in Fig. 6) the overall performance of FAIR-1 is better compared to FAIR-0 because of the better discriminability (red curve). Overall, FAIR-1 is significantly less sensitive to noise than FAIR-0. Greater sensitivity of SIFT-O to noise is due to the loss of discriminability by rotational normalization. Since image values are scaled to the interval of $[0, 1]$, the maximum noise corresponds to SNR=20dB.

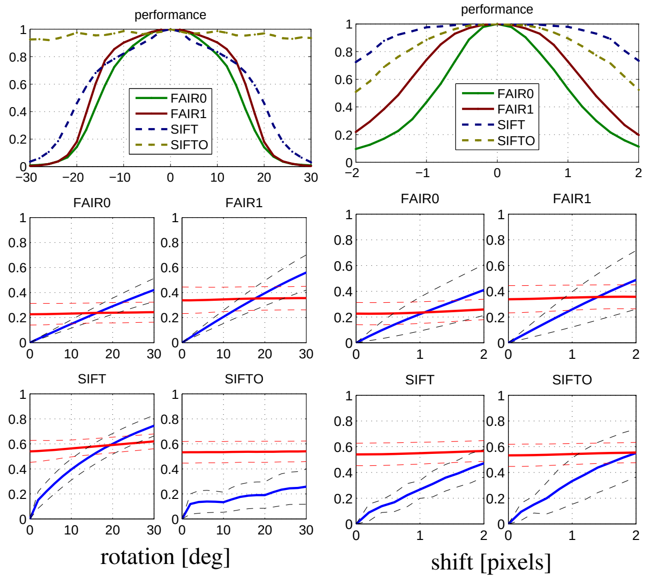

Fig. 4. Performance under rotation (left) and shift (right).

4 Conclusions and Future Work

In this paper we have proposed a local image neighborhood descriptor suitable for matching under affine distortion. The goal was not to achieve total affine invariance of the descriptor itself but rather reasonable insensitivity to residual affine transformation after affine adaptation process. Affine aggregation region is a necessary component of any finite affine descriptor since when comparing two such descriptors one must be able to compare the same image measurements. Therefore good behavior means that the descriptor degrades slowly for affine mappings near identity. The goal of such slow degradation has been reached as has been demonstrated by the experiments. For a good recognition performance, discriminability is a more critical parameter. A suitable method for direct comparison of discriminabilities between descriptors of different dimensions remains an open problem.

FAIR has significantly smaller dimension (30) than SIFT (128). Small representation allows us to use more elaborate matching method, as has been hinted in Sec. 2.

Localization of FAIR in scale space is about twice better than in SIFT which allows for greater accuracy when used for structure from motion problems.

From the computational complexity point of view, the most expensive part of FAIR is image interpolation needed to collect raw measurements m_0 or m_1. The interpolation should be at least linear. This is similar to SIFT. The means and standard deviations needed to construct the descriptor are not expensive, one could consider replacing the standard deviation of m_1 with the mean of m_0 for speeding the preprocessing up.

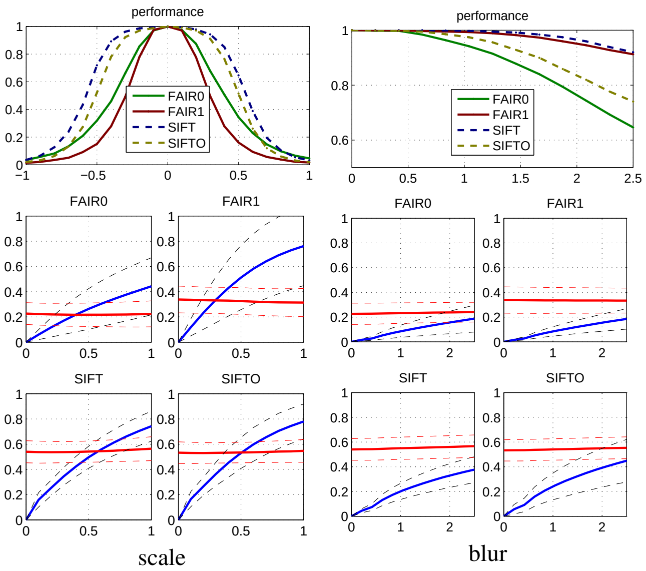

Fig. 5. Performance under scale (left) and blur (right).

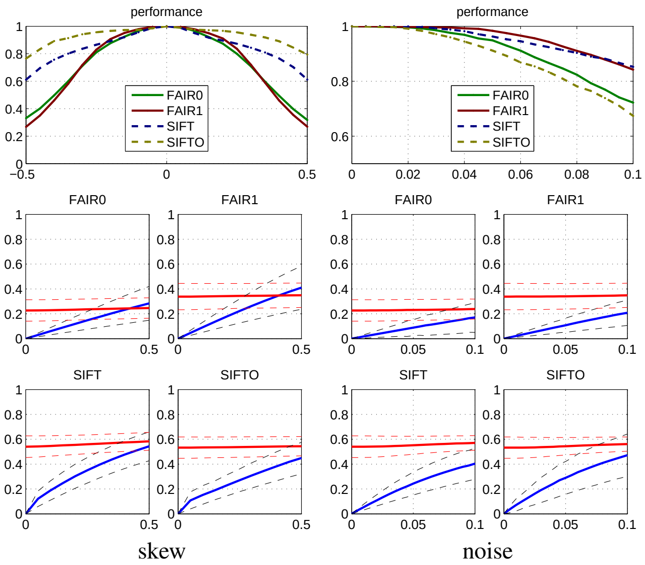

Fig. 6. Performance under skew (left) and noise (right).

In our future work we will try maximizing the performance of FAIR by trying alternative neighborhood sampling schemes, by varying the shape of the neighborhood and its other parameters like sampling rate, sector width, and the angular step. We will find the optimal dimension of the feature, the dimension used here was chosen arbitrarily. The dynamic programming warping method should be very fast and is also a topic for further work. We also plan experimenting with the use of additional statistics describing the individual segment histograms.

Acknowledgments We thank David G. Lowe for making it possible to use his implementation of SIFT and to Michal Perd'och for his interface of SIFT to Matlab. This work has been supported in part by The Czech Ministry of Education under project 1M0567, in part by The Czech Academy of Sciences under project 1ET101210406 and by the EU Grant FP6-IST-027113 eTRIMS.

A Invariance of m_1

Let $f(\mathbf{x})$ be the original image and $g(\mathbf{A}\mathbf{x})$ be the image with its domain affinely distorted so that

$$g(\mathbf{A}\mathbf{x}) = f(\mathbf{x}). \tag{5}$$

We show the measurement m_1 defined by (1) is invariant to $\mathbf{A}$, i.e. it holds $\mathbf{y}^\top \nabla_\mathbf{y}\, g(\mathbf{y}) = \mathbf{x}^\top \nabla_\mathbf{x} f(\mathbf{x})$ if $\mathbf{y} = \mathbf{A}\mathbf{x}$ is the domain transformation, where we have abbreviated $\nabla_\mathbf{x} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})$. Applying the gradient operator to both sides of (5) we have

$$\nabla_\mathbf{x}\, g(\mathbf{A}\mathbf{x}) = \nabla_\mathbf{x} f(\mathbf{x}). \tag{6}$$

By the chain rule on the left side of (6), we get $\nabla_\mathbf{x}\, g(\mathbf{y}) = \mathbf{A}^\top \nabla_\mathbf{y} g(\mathbf{y})$ which transforms (6) to

$$\nabla_\mathbf{y}\, g(\mathbf{y}) = \mathbf{A}^{-\top} \nabla_\mathbf{x} f(\mathbf{x}). \tag{7}$$

We can then write $\mathbf{y}^\top \nabla_\mathbf{y}\, g(\mathbf{y}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}^{-\top} \nabla_\mathbf{x} f(\mathbf{x}) = \mathbf{x}^\top \nabla_\mathbf{x} f(\mathbf{x})$, QED.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans PAMI*, 24(4):509–522, 2002.

[2] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proc CVPR*, pages 511–517, 2004.

[3] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using affine-invariant regions. In *Proc CVPR*, pages 319–324, 2003.

[4] T. Lindeberg and J. Gårding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int J Computer Vision*, 60(2):91–110, 2004.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[7] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans PAMI*, 27(10):1615–1630, 2005.