Geometric Hashing with Local Affine Frames

Ondřej Chum[a] and Jiří Matas[a,b]
[a]Center for Machine Perception, Czech Technical University in Prague, Czech Republic
[b]CVSSP, University of Surrey, Guildford, UK
[chum,matas]@cmp.felk.cvut.cz

Abstract

*We propose a novel representation of local image structure and a matching scheme that are insensitive to a wide range of appearance changes. The representation is a collection of local affine frames that are constructed on outer boundaries of maximally stable extremal regions (*MSERs*) in an affine-covariant way. Each local affine frame is described by a relative location of other local affine frames in its neighborhood. The image is thus represented by quantities that depend only on the location of the boundaries of* MSERs*. Inter-image correspondences between local affine frames are formed in constant time by geometric hashing. Direct detection of local affine frames removes the requirement of point-based hashing to establish reference frames in a combinatorial way, which has in the case of affine transform complexity that is cubic in the number of points. Local affine frames, which are also the quantities represented in the hash table, occupy a 6D space and hence data collisions are less likely compared with 2D point hashing.*

Experimentally, the robustness of the method and its insensitivity to photometric changes is demonstrated on images from different spectral bands of satellite sensor, on images of a transparent object and on images of an object taken during day and night.

1. Introduction

Methods[1] based on matching of regions repeatably detected in a transformation-covariant manner have demonstrated impressive object recognition capabilities. Noteworthy applications include Šivic's and Zisserman's Video-Google system capable of object retrieval in feature-length films [17] and Lowe's real-time object recognition system [8]. The impact of the methodology is clear from the range of computer vision problems it has been successfully applied to: wide baseline stereo matching [1, 9, 13, 21, 19], image retrieval from large databases [15, 20], model based recognition [8, 10, 14], object retrieval in video [17], texture recognition [7], robot localization [16], and panoramas [2].

Figure 1. Detecting a logo (central image). Successful detection in images 1-8 is highlighted by a quadrilateral. Note the differences in appearance. Bottom row: cut-outs from images 6, 7, 8, and 3.

All the listed approaches rely on establishing correspondences of local image patches that are projections of the same pre-image[2], *i.e.* whose appearance in the matched images is identical, modulo affine local geometric deformation and affine photometric transformation. In the paper, we address a more general situation where *it is only assumed that intensity discontinuities are preserved*; no other assumptions about photometric changes or object appearance are needed. Such situation is common *e.g.* in the case of matching images from different modalities or spectral bands as in medicine or remote sensing (see Fig. 4) or when matching

[1]The authors were supported by EU project MRTN-CT-2004-005439 VISIONTRAIN, by EU project IST-004176 COSPAL, and by The Czech Ministry of Education project 1M0567 CAK.

[2]With the exception of texture recognition where the statement is true only in a statistical sense.

images taken during day and night. Consider a different instance of the situation: the nine images depicted in the top part of Fig. 1. A company logo appears in all of the images in a number of different color variations, but the position of discontinuities carry information sufficient for its recognition.

We propose a novel image descriptor that is insensitive to a range of appearance changes. The image is represented by a collection of local affine frames (LAFs) [10] that are constructed on outer boundaries of maximally stable extremal regions (MSERs) [9] in an affine-covariant way. The description of a LAF needed for establishing tentative correspondences is provided by the relative location and pose of other LAFs in its neighborhood. The neighborhood is also defined in an affine-covariant manner. As a result, the image representation depends only on MSER boundaries and the proposed method successfully matches two views under the condition that the boundaries have sufficient repeatability. It is shown experimentally that the condition is commonly satisfied. The choice of MSERs is not critical for the method and any affine-covariant detector that permits local affine frame constructions could be used. MSERs were chosen since the detector is available[3] and it performed well in a recent comparison paper [11].

As a second contribution, we propose a geometric hashing [6] method that establishes each tentative correspondence between local affine frames in constant time. We show that local affine frames fit well into the geometric hashing framework, preserving all the advantages of the method and yet overcoming its limitations. In the original Lamdan and Wolfson method [6], for the case of affine transformation between images, N 2D points are invariantly represented in $\mathcal{O}(N^3)$ affine frames formed from all triplets of points. The resulting hash table thus has $\mathcal{O}(N^4)$ entries. Since each entry is described by a 2D position, data collision is likely. Direct detection of local affine frames removes the $O(N^3)$ factor. Local affine frames (which are also the objects stored in the hash table) occupy a 6D space and hence data collisions are unlikely compared with 2D point hashing.

Note that geometric hashing introduces robustness to the process of establishing tentative correspondences. This means that (i) the evidence for a correspondence can originate in any part of the image and (ii) it does not matter if the local affine frame is near a 3D discontinuity, where some fraction of frames in its neighborhood will vote for a different affine image-to-image transformation. Compare this with the standard SIFT [8] descriptor matching. If the SIFT is computed from an area whose pre-image is straddling a 3D discontinuity, the descriptor is unlikely to match after changing a viewpoint[4]. The robustness of the LAF matching process is demonstrated in an experiment where a partially transparent object is matched on completely different backgrounds. The problem of wide-baseline image matching in the presence of severe intensity changes has been address by the Dual-bootstrap ICP approach of Stewart *et al.* [18]. The method presented in this paper exploits hashing (is non-iterative) and relies on the success of the affine-covariant region detector, [18] is an iterative method. The idea of using a pair of features rather than a single feature has also been used in the work of Tell and Carlson [19].

The rest of the paper is structured as follows. First, definitions used throughout the paper are introduced in Section 2. The details of the matching method based on the geometric hashing are laid out in Section 3. Experimental validation follows in Section 4 and the paper is concluded in Section 5.

2. Definitions

By a LAF we understand an ordered triplet of non-collinear points $L = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, where $\mathbf{x}_i = (x, y, 1)^\top$. Let $L^{\llcorner} = (\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3)$, where $\mathbf{n}_1 = (1, 0, 1)^\top$, $\mathbf{n}_2 = (0, 0, 1)^\top$, and $\mathbf{n}_3 = (0, 1, 1)^\top$, be a canonical LAF. Let normalization $\mathtt{N}$ be an affine transformation that transforms LAF L to a canonical frame $L^{\llcorner} = \mathtt{N}L$. Let $\mathtt{A}$ be a matrix representing an affine transformation with the last row $(0, 0, 1)$; and let $\mathtt{UDV}^\top$ be a SVD decomposition of the upper left 2×2 submatrix of $\mathtt{A}$. Let $\mathtt{D} = \mathrm{diag}(d_1, d_2)$, where $d_1 \geq d_2$. We define *anisotropy factor* of an affine transformation $\mathtt{A}$ as $a(\mathtt{A}) = d_1/d_2$. The anisotropy factor of LAF is the anisotropy factor of its normalization.

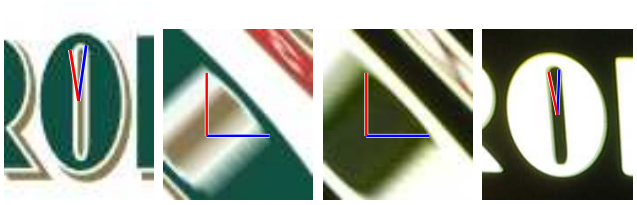
3. Geometric Hashing with LAFs

The process of establishing tentative correspondences between two images selects pairs of potentially corresponding image elements. The image elements considered here are local affine frames. A standard approach in wide-baseline stereo matching algorithms is to describe each image element by an affine invariant descriptor. Affine invariance is chosen because the perspective projection of a close-to-planar surface is locally well approximated by an affine transformation. Tentative correspondences are formed on the basis of descriptor similarity. The invariant measurements are (functions of) the image intensities within some affinely covariantly defined shape (ellipse, parallelogram), *e.g.* [21, 9].

The affine invariant descriptor constructed in this paper is derived from the mutual position of two LAFs: one, called *reference frame* (RF), provides a coordinate system; the other, called *description frame* (DF), provides a

[3] Executable of the MSER detector is available at `http://www.robots.ox.ac.uk/~vgg/research/affine`.

[4] Unless the background is uniform.



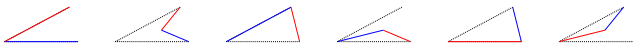
L $L^{\perp}$ $L^{\perp\prime}$ L'

Figure 2. Normalization of a LAF is not only dependent on the viewpoint, but mainly on the the shape of the LAF itself (crops from Fig. 1 - logo image and image 3). The anisotropy factor of the normalizations are 6.7 and 8.0 respectively, while the anisotropy factor of the affine transformation mapping L to L' is 1.3.

6-dimensional affine invariant vector (two dimensions for each of its three points). We exploit the idea of geometric hashing [6]: discretized 6D descriptors are used as hashing keys, LAFs with similar descriptors are thus found (and tentative correspondences formed) in constant time.

Reference frame construction and selection. We assume that uncertainty regions of locations of points forming LAFs are circles (*i.e.* assuming isotropic Gaussian distribution of positional error). The circles are transformed to elliptical uncertainty regions in the normalized frame. A reference frame with high value of anisotropy factor (of its normalization transformation) leads to elongated ellipses of uncertainty that cover a number of bins (independently of the choice of the discretization). Consider, for example, LAFs where the axis form an angle of almost 0 or π. The coordinate systems defined by such LAFs are significantly deformed. In general, no restrictions on the anisotropy factor of RFs can be used. Nevertheless, in wide baseline stereo matching, assumptions on the affine transformation are often made. Wide range of off-plane rotation of the viewpoint results in a similar anisotropy factor of the image transformation [8]. Also, high anisotropy together with image discretization often causes failure in the first step of the matching process, *i.e.* in the detection of affine-covariant regions.

In many cases, the anisotropy factor of a LAF is significantly higher than the anisotropy factor of the transformation between matched images, see the example in Fig. 2. We propose to use more than a single reference frame associated with a LAF. The restrictions on the image transformation can be then applied to the RFs. For wide-baseline stereo matching, we propose six RF constructions, depicted in Fig. 3. Then, there are six hashing tables for each type of RF construction, *i.e.* descriptors originated from the same type of RF construction can be matched. Not all six RFs are used for every frame, since the RFs with high value of the anisotropy factor are more sensitive to noise. Only those RFs having the anisotropy factor smaller than a threshold, in our experiments set to the value of 4, are used for affine invariant description.



4.00 1.33 2.16 3.42 2.16 3.42

Figure 3. Different construction of coordinate frames (some involving the center of gravity) and the anisotropy factor of the corresponding normalization.

Selection of description LAFs. In order to select description LAFs in an affine invariant fashion, the selection has to be based on affine invariant measurements. Since the affine transformation is only a locally valid approximation of the real image transformation, the description LAFs are selected in proximity of the reference frame (in our case measured as a distance from the origin of the reference frame in the RF coordinate system). If there were no other selection criteria, large RFs would use a large number of (even all) frames as description LAFs (leading to a computational explosion). In our approach, the following additional criterion is used: the reference frame and the description LAF are of similar scale. Such approach is affine invariant, and guarantees (under certain reasonable assumptions on the feature detector) an independence of the number of description LAFs of the size of the RF. Moreover, if a feature is detected in both images with a change of scale, it is likely that features with a similar scale will survive the image transformation too.

Lazebnik *et al.* [7] suggested to select description points that are closer than certain number of pixels (measured in the coordinate system of the image). However, this approach is not affine-covariant and is very sensitive to scale changes.

Choice of the description space and its discretization. There is a number of possible representations of the 6D affine invariant descriptor. Let L_1, L_2 be two LAFs in the first image and L_1', L_2' in the second image. We express points of L_2 in a coordinate system derived from L_1. In our implementation, we have chosen the coordinates of the 6D descriptor as polar coordinates of points of the description LAF L_2. The first two dimensions are set as polar coordinates of the central point of the description LAF, having the origin of the coordinate system at the origin of the RF. The angle is discretized into 25 bins and the distance into 16 bins. The remaining four dimensions are given by the polar coordinates of the two remaining points from the description LAF. The origin of the polar coordinate system is located at the central point of the description LAF. The angles are discretized into 25 and the radii into 6 bins. This gives $9 \cdot 10^6$ possible values of the descriptor.

Votes counting. The votes are counted in a sparse matrix represented as a hashing table [5]. The collisions were handled by a secondary hashing function. Even if the same pairs of LAFs appear in an identical bin for more than a single construction of reference frame the vote is counted only once.

4. Experiments

The proposed approach was tested on two-view estimation of homographies and, in one case, epipolar geometry. We focused on homographies, since in this case (i) ground truth is easily established and (ii) the induced one-to-one correspondence of pixels is easy to visualize. The detector of MSER regions [9], which was used in all experiments with default parameters, outputs two disjoint sets of extremal regions: MSER+ containing regions with the inside brighter than outside and MSER- with opposite contrast. In both [9] and [11], tentative correspondences of MSER+ and MSER- were established separately, *i.e.* the positive and negative contrast regions were not allowed to match. Since one of our objectives is the ability to match under arbitrary changes of illumination and/or contrast, a union of MSER+ and MSER- was formed. The LAFs were then constructed on all regions following the procedures described in [10].

Where is the logo? In the experiment, a logo (Fig. 1-center) was sought in the other eight images. To localize the logo in the test images, a homography [4] was robustly estimated by RANSAC [3]. The logo image was downloaded from the internet and has 450×251 pixel resolution, test images were shot by a four megapixel digital camera (size 2272×1707). As shown in (Fig. 1), in all cases the logo was correctly detected and localized. The images were chosen to highlight the potential of a method exploiting only locations of discontinuities. Image 2 (top, center) shows a sticker-on-glass version of the logo. Somewhat reminiscent of M.C. Esher's "Three Worlds", the image is a superposition of object behind the glass and reflected from the glass. Images 4,5,6 and 7 show the same object during day and night; Image 4 was taken with a flash. Cut-outs at the bottom of Fig. 1 demonstrate the variability of appearance of the logo in images 6, 7, 8 and 3. We mention in passing that the scale of the logo changes from 0.5 of the original (test image 9) up to 4.9 (test image 2). The current (suboptimal Matlab) implementation outputs tentative correspondences for an image pair in terms of seconds.

The matching algorithm was also successfully applied to all $\binom{9}{2}$ image pairs.

Inter-spectral matching is demonstrated on images acquired by the Landsat Thematic Mapper [12]. The images shown in Fig. 4 are the thermal (left) and the mid infra-red band (right) respectively. The thermal and infra-red sensors have different resolution. Registration information is available; in the experiment we pretend it is unknown and try to estimate it as a homography. Note the differences in appearance (intensity) of the two images that are due to changes of spectral reflectances of individual materials. Such changes do not follow a simple model and their effects cannot be removed by normalization. Nevertheless, location of discontinuities is often preserved; transitions between materials

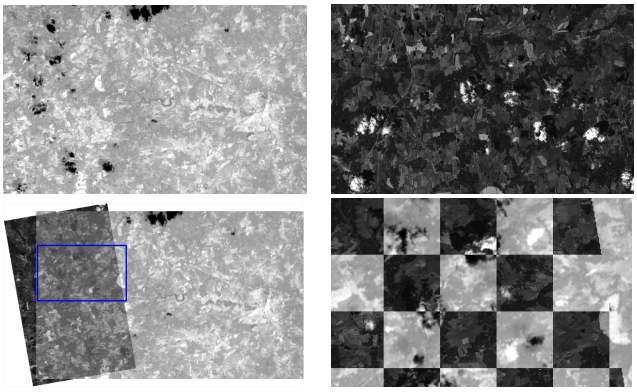


Figure 4. Matching spectral bands of Landsat TM images (top row). Spatial resolution of the thermal band image (left, 10.40 – 12.50 micrometer range) is half the resolution of the mid-infrared image (right, 2.08 – 2.35 micrometer range). Registered images or shown as a mosaic (bottom left) and checkerboard overlay (bottom right). Note the contract reversal in parts of the image.

(e.g. vegetation type) typically implies step-edge response in both bands. The variation of appearance, including contrast reversal, can be visually observed in Fig. 4 (bottom right) on the checkerboard overlay. The bands were successfully matched; a homography consistent with 49 LAF correspondences was found by RANSAC. The quality of tentative correspondences, measured by the density of inliers, was high – 38% in the 100 top-ranked correspondences.

Graffiti experiment. Results on the Graffiti set[5] have been reported in a number of publications (*e.g.* [11] and the references within). We include an experiment on one of the most difficult pairs (see Fig. 5a) to allow comparison. In this case, appearance-based methods work well - there is no visible change in illumination. Ignoring color and intensity information is only making the matching problem more challenging. There were 2337 LAFs (818 from MSER+ and 1519 from MSER- respectively) in the first image (Fig. 5a) and 4325 (1735 + 2590) LAFs in the second image(Fig. 5b) output by the detector. The pair of images was successfully matched.

We first focus on the quality of tentative correspondences generated by geometric hashing. Since the ground truth is known, it can be established that 88 LAF were detected in both images (*i.e.* there is a corresponding LAF in the other image with Sampson's error [4] under 2 pixels). If tentative correspondences were sought assuming an arbitrary change of the intensities, *i.e.* LAFs originating from MSER+ and MSER- were allowed to match, then the total number of correct correspondences is 64. The dark solid curve plotted in Fig. 6 shows the fraction of inliers among n top-ranked tentative correspondences (ranked by the number of votes). The density of inliers is such that RANSAC selects a solution

[5]The Graffiti images are available at
http://lear.inrialpes.fr/people/Mikolajczyk/Database/graff6.tar.gz

(a) (b) (c)

Figure 5. The Graffiti experiment. Left image (a) with superimposed 88 LAFs that were detected in both images. Right image (b) with superimposed boundary of the transformed left image . (c) The anisotropy factor of an affine transformation locally approximating the ground truth homography of the left-to-right image mapping.

in only a few iterations.

We next assess how much is gained if we exploit the *a priori* knowledge that illumination did not change. In this case, LAFs originating from MSER+ and MSER- were not allowed to match. The number of correctly matched correspondences increased to 80, which is close to the possible maximum of 88. The light solid curve plotted in Fig. 6 shows the fraction of inliers among n top-ranked tentative correspondences (ranked by the number of votes) when MSER+ and MSER- were matched independently. Finally, we assess the benefit of defining multiple reference frames for each LAF. The dashed curve of Fig. 6 shows the fraction of inliers among n top-ranked tentative correspondences if only a single reference frame is used; MSER+ and MSER- were joined in this experiment. Only 32 correct correspondences were among the tentative correspondences formed. The introduction of multiple reference frames thus doubled both the number and density of correct tentative correspondences.

Plot (Fig. 5c) shows the spatial distribution of the anisotropy factor of an affine transformation locally approximating the projective transformation of the images. The values are plotted for pixels of the left image that are mapped into (have an image in) the right image.

The ground truth homography $\mathtt{H}$ was approximated in each point $\mathbf{x}' = \mathtt{H}\mathbf{x}$ by a first order approximation (affine transformation $\mathtt{A}$)

$$\mathtt{A} = \begin{pmatrix} h_1 - x'h_7 & h_2 - x'h_8 & xx'h_7 + yx'h_8 + h_3 \\ h_4 - y'h_7 & h_5 - y'h_8 & xy'h_7 + yy'h_8 + h_6 \\ 0 & 0 & h_7x + h_8y + h_9 \end{pmatrix}.$$

The anisotropy factor $a(\mathtt{A}(\mathtt{H}, \mathbf{x}))$ reaches 4.3 in the rightmost part of the image, while the threshold on the anisotropy factor of the normalization transformation $a(\mathtt{N})$ of coordinate frames was set to 4. Note, that this *does not* mean, that LAFs from this region cannot be matched. It only means that coordinate frames with the anisotropy factor $a(\mathtt{N})$ below 1.075 have no chance of being matched.

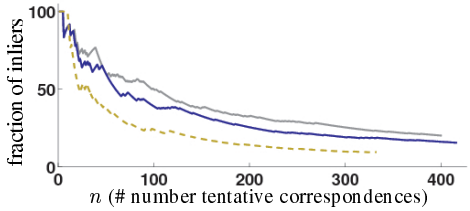


Figure 6. The fraction of inliers among n top-ranked tentative correspondences. Correspondences are ranked by the number of votes in the hash table.

Figure 7. Two photographs of a text on a transparent foil, each time on a different background, and their superposition after successful registration.

Arbitrary background - text on a transparency. A text printed on a transparent foil was captured on two different backgrounds. Some of the letters are detected in both images. An affine-covariant measurement region larger than the detected letter includes the background, which is never the same in the corresponding parts of the two images. A measurement region that covers the letter (or its part) only, is not discriminative, as most of the letter appears many times in the text. On the other hand, groups of letters and their mutual positions provide enough information for the matching – the two images were successfully registered, see Fig. 7 (right).

Epipolar geometry estimation. This experiment demonstrates the performance of the proposed method on a non-planar scene. The correct epipolar geometry was recovered

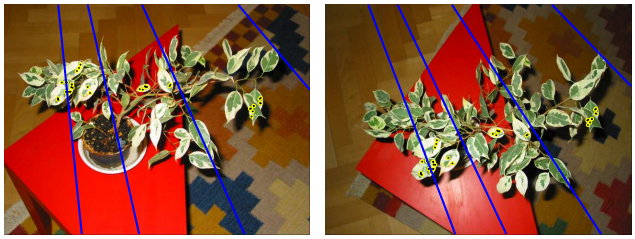


Figure 8. The plant scene. An image pair with 50 superimposed correspondences consistent with the epipolar geometry.

by RANSAC using tentative correspondences obtained by the proposed method.

Comparison with MSER-LAF approach [10]. On images with color (intensity) preserved (Figs. 5, 8), both algorithms output approximately equal number of correspondences with the same percentage of inliers. In experiment in Figs. 4, 7 and on most image pairs from Fig. 1 the method [10] failed.

5. Conclusions

We proposed a novel image representation – a collection of local affine frames that are constructed on outer boundaries of maximally stable extremal regions in an affine-covariant way. We showed how inter-image correspondences between LAFs can be formed in constant time by geometric hashing. Since LAFs are directly detected in images, there is no need to establish reference frames combinatorially, which leads to very significant efficiency gains. We showed that due to noise, the choice of representation of a reference frame on a LAFs is an important technicality.

Experimentally, the performance of the method was demonstrated on two-view matching problems of images from different modalities (multi-spectral images), images of a transparent object on variable backgrounds and on images where albedo changed arbitrarily.

The method was show to perform comparably well to an orthogonal method [10] that uses patch appearance only to establish the correspondences. For the matching tasks where the appearance is preserved, the two methods can be combined.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *CVPR00*, pages I:774–781, 2000.

[2] M. Brown and D. Lowe. Recognising panoramas. In *Proc. ICCV03*, volume I, pages 1218–1225, October 2003.

[3] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, June 1981.

[4] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University, Cambridge, 2nd edition, 2003.

[5] D. E. Knuth. *The Art of Computer Programming : Sorting and Searching*, volume 3. Addison Wesley, Boston, USA, 2nd edition, 1998.

[6] Y. Lamdan and H. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proc. ICCV*, pages 238 – 249, 1988.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *ICCV'03*, pages 649–655, October 2003.

[8] D. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

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

[10] J. Matas, Š. Obdržálek, and O. Chum. Local affine frames for wide-baseline stereo. In *Proc. ICPR*, volume 4, pages 363–366. IEEE CS, Aug 2002.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1/2):43–72, 2005.

[12] NASA Landsat Program. Landsat. U.S. Geological Survey. http://www.landcover.org.

[13] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. ICCV*, pages 754–760, 1998.

[14] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. CVPR*, volume II, pages 272–277, 2003.

[15] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, May 1997.

[16] S. Se, D. Lowe, and J. Little. Local and global localization for mobile robots using visual landmarks. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 414–420, Maui, Hawaii, Oct 2001.

[17] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, pages 1470 – 1477, Oct. 2003.

[18] C. V. Stewart, C.-L. Tsai, and B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Trans. on Medical Imaging*, 22(11):1379–1394, Nov 2003.

[19] D. Tell and S. Carlsson. Cobmining appearance and topology for wide baseline matching. In *Proc. 7th ECCV*, volume 1, pages 68–81. Springer-Verlag, 2002.

[20] T. Tuytelaars and L. V. Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc 3rd Int'l Conf. on Visual Information Systems*, pages 493–500, 1999.

[21] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *Proc. 11th BMVC*, 2000.