International Conference on Pattern Recognition (ICPR2002)
August 2002, Quebec City, Canada

Local Affine Frames for Wide-Baseline Stereo

Jiří Matas[1,2], Štěpán Obdržálek[1], Ondřej Chum[1]
[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 procedure for establishing wide-baseline correspondence is introduced. Tentative correspondences are established by matching photometrically normalised colour measurements represented in a local affine frame. The affine frames are obtained by a number of affine invariant constructions on robustly detected maximally stable extremal regions of data-dependent shape. Several processes for local affine frame construction are proposed and proved affine covariant. The potential of the proposed approach is demonstrated on demanding wide-baseline matching problems. Correspondence between two views taken from different viewpoints and camera orientations as well as at very different scales is reliably established. For the scale change present (a factor more than 3), the zoomed-in image covers less than 10% of the wider view.

1. Introduction

Establishing reliable correspondences between a pair of images taken from arbitrary viewpoints is a fundamental problem in many computer vision tasks. Applications include 3D scene reconstruction, motion recovery, image mosaicing, content-based image retrieval, mobile robot navigation and many more. In the wide-baseline set-up, local image deformation cannot be realistically approximated by a translation or a translation with rotation, and a full affine model is required. Correspondence cannot be therefore established by comparing regions of a fixed shape, like rectangles or circles, since their shape is not preserved under the group of transformations that occur between the images.

In the literature, correspondences have been traditionally sought by matching features computed on local neighborhoods of detected interest points [11, 7, 1, 9]. In this paper, we rely on the so-called Maximally Stable Extremal Regions introduced in [2], which were shown to be highly repeatable over a wide range of image formation conditions. Local affine frames are established on the regions by several affine invariant constructions. The approach obtains region correspondences in a robust manner, since multiple local frames are associated with each region. The proposed approach departs from standard methods based on computing invariants [7, 3]. Fully affine-invariant descriptors were introduced recently, exploiting local texture characteristics [1], or local configuration of multiple image edges or interest points [5, 10]. Schaffalitzky and Zisserman [6] presented a method based om automatic determination of local neighborhood shape, but only for image areas where stationary texture occurs.

The main contribution of the paper is the utilization of several processes for affine-invariant determination of local affine frames (local coordinate systems) on distinguished regions of a complex shape. Robustness of the matching procedure is accomplished assigning multiple frames to each region, and not requiring all of the frames to match. Should, for example, a region be partially occluded in one of the images, only a subset of the associated frames will correspond. Using measurements on these frames, we are able to successfully solve non-trivial instances of the problem of establishing inter-image correspondences. We experimentally show that the measurements are sufficiently stable even in the presence of substantial scale change and other affine deformations.

The paper is organized as follows. In Section 2 we briefly review the concept of distinguished regions. Section 3 gives a description of procedures giving local affine frames on distinguished regions. Section 4 details how tentative 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

*The authors were supported by the EU project IST-2001-32184, the Grant Agency of the Czech Republic project GACR 102/00/1679 and CTU grant No. CTU0209613.

over a range of image formation conditions. In this work we exploit a new type of distinguished regions introduced in [2], 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 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 [2] which includes a formal definition of the MSERs and a detailed description of the extraction algorithm. The report [2] is available online.

3. Local Frames of Reference

In the literature on wide-baseline stereo, invariants have been widely used. In the presented method, invariants would be used only in a preliminary selection of potential correspondences, i.e. for efficiency reasons. Instead, tentative correspondences are established by matching photometrically normalised colour measurements represented in a local affine frame. Our experiments show that such approach produces a higher percentage of correct correspondes. The frames are obtained by a number of affine invariant constructions. The constructions exploit properties of the center of gravity, second order algebraic moments, convex hull, and properties of sets of parallel lines. It is not required that a local affine frame be constructed for every distinguished region. Indeed, no direction can be preferred 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 constructed in a stable and thus repeatable way. Robustness of the approach is thus achieved by selecting only stable frames and employing multiple processes for frame computation. In particular, the following processes are applied to each distinguished region:

- Region normalisation by the covariance matrix followed by the detection of extremal points with respect to the center of gravity
- Detection of stable bi-tangents and of points of maximal distance from the bi-tangents.

Invarinace of the bi-tangets is a consequence of the affine invariance (and even projective invariance) of the convex hull construction [8, 4].

Two types of frames are derived from the bi-tangants. In the first type, the center of gravity is the third point completing the affine frame. The second type uses the point in the corresponding concavity most distant from the bi-tangent line. Affine covariance of the center of mass can be shown trivially. The covariance of the point of maximal distance from a line is easily appreciated taking into account that affine transform maintains parallelism of lines and their ordering.

In the rest of the section we show the invariance of the construction using the covariance matrix. An 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. Let's consider a region Ω_1, and its image Ω_2. Area of Ω_2 is given as

$$|\Omega_2| = \int_{\Omega_2} d\Omega_2 = \int_{\Omega_1} |A|\, d\Omega_1 = |A||\Omega_1|, \tag{1}$$

The center of gravity of region Ω is $\mu = \frac{1}{|\Omega|}\int_\Omega \mathbf{x} d\Omega$ where $\mathbf{x} = (x, y)^T$. The relation between the centers of gravity of transformed regions is:

$$\begin{aligned} \mu_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2} \mathbf{x_2}\, d\Omega_2 = \frac{1}{|A||\Omega_1|}\int_{\Omega_1} (A^T\mathbf{x_1}+\mathbf{t})|A|\, d\Omega_1 \\ &= A^T\frac{1}{|\Omega_1|}\int_{\Omega_1}\mathbf{x_1}\, d\Omega_1 + \frac{1}{|\Omega_1|}\int_{\Omega_1}\mathbf{t}\, d\Omega_1 \\ &= A^T\mu_1 + \mathbf{t} \end{aligned} \tag{2}$$

so the center of gravity changes covariantly with the affine transform. The covariance matrix Σ of a region Ω is a 2x2 matrix defined as $\Sigma = \frac{1}{|\Omega|}\int_\Omega (\mathbf{x}-\mu)(\mathbf{x}-\mu)^T d\Omega$. Covariance matrix of a transformed region Ω_2 is then

$$\begin{aligned} \Sigma_2 &= \frac{1}{|\Omega_2|}\int_{\Omega_2}(\mathbf{x_2}-\mu_2)(\mathbf{x_2}-\mu_2)^T d\Omega_2 \\ &= \frac{1}{|A||\Omega_1|}\int_{\Omega_1}(A^T\mathbf{x_1}+\mathbf{t}-(A^T\mu_1+\mathbf{t})) \\ &\quad (A^T\mathbf{x_1}+\mathbf{t}-(A^T\mu_1+\mathbf{t}))^T|A|\, d\Omega_1 \\ &= \frac{1}{|\Omega_1|}\int_{\Omega_1}(A^T(\mathbf{x_1}-\mu_1))(A^T(\mathbf{x_1}-\mu_1))^T d\Omega_1 \\ &= A^T\left(\frac{1}{|\Omega_1|}\int_{\Omega_1}(\mathbf{x_1}-\mu_1)(\mathbf{x_1}-\mu_1)^T d\Omega_1\right)A \\ &= A^T\Sigma_1 A \end{aligned} \tag{3}$$

Cholesky decomposition of a symmetric and positive-definite matrix Σ is a factorization $\Sigma = U^TU$, where U is an upper triangular matrix. Cholesky decomposition is defined up to a rotation, since $U^TU = U^TR^TRU$ for any rotation R. For the decomposition of covariance matrix of a transformed region we write

$$\Sigma_2 = U_2^TR_2^TR_2U_2 = A^T\Sigma_1 A = A^TU_1^TR_1^TR_1U_1A$$

thus

$$R_2U_2 = R_1U_1A \tag{4}$$

$$U_2 = R_2^{-1}R_1U_1A = RU_1A \tag{5}$$

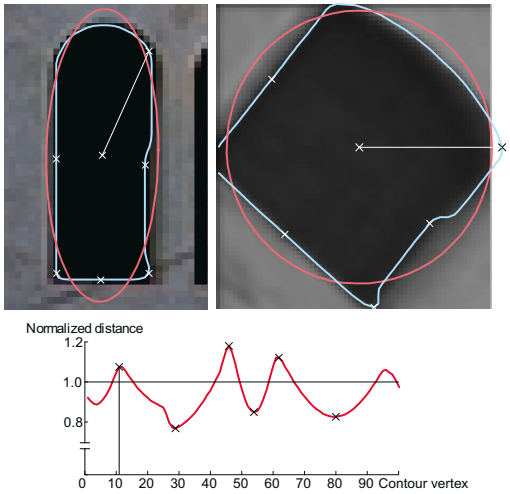

Figure 1. Construction of affine frames.

Hence the triangular matrix U, obtained through the cholesky-decomposition of a covariance matrix Σ, is covariant, up to an arbitrary orthonormal matrix R, with the affine transform applied to the region.

The process of extremal point detection is depicted in Figure 1. A region detected in an image (at top left) is transformed to the shape-normalized frame (top right). Normalized distances of contour points are plotted at the bottom. The ellipse defined by the covariance matrix of the region is transformed to the unit circle in the normalized frame.

4. Establishing tentative correspondences

We wish to solve the stereo problem in the correspondence space. The first two steps of the process, detection of distinguished regions and construction of local affine frames has been described above. The task now is to establish tentative correspondences between the two images. Since full affine frames are known, invariant descriptors of local appearance are not needed. The selection of tentative correspondence can rely simply on correlating intensity-normalised regions defined intrinsically in terms of the local coordinate frames.

Having two sets of local affine frames, set S_1 of frames computed on the regions of the first image, and set S_2 on the second image, of the matching procedure can be outlined as follows:

1. For all frame pairs ($f_1 \in S_1, f_2 \in S_2$), so that f_1 and f_2 are of the same type do:

2. Calculate intensity-normalised cross-correlation c between f_1 and f_2.

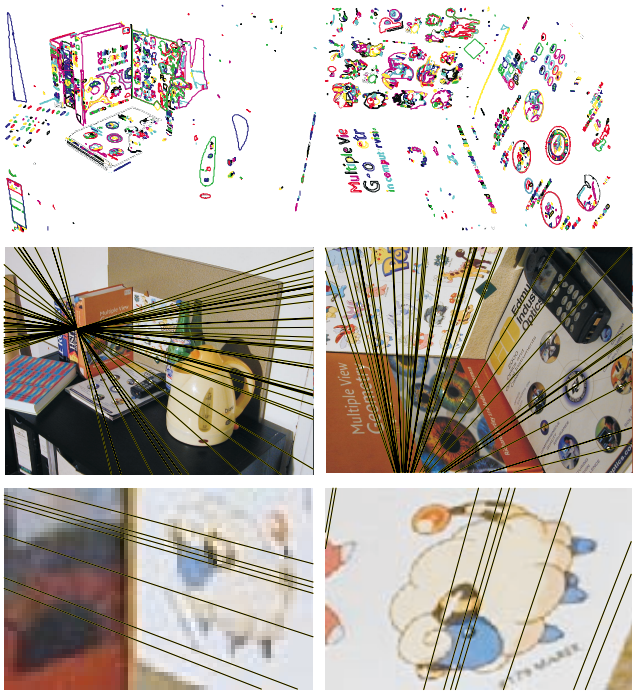

Figure 2. Bookshelf. An image pair with significant scale ($3.3\times$), rotation and skew.

	Bookshelf		Valbonne	
	left	right	left	right
n of regions	1313	1764	1154	877
n of local frames	4482	6622	3656	2726
n of tent. corresp.	109		54	
EG consistent	42		23	

Table 1. Summary of the matching process.

3. Reject a frame pair, if there is a frame $f_2' \in S_2 \setminus f_2$ or a frame $f_1' \in S_1 \setminus f_1$ such that corr$(f_1, f_2') > c$, or corr$(f_1', f_2) > c$ (keep only mutually closest frames).

4. A pair of regions forms a tentative correspondence if either one pair of their frames matched with high correlation or multiple frame pairs matched.

5. Experiments

In the experiments, *Maximally Stable Extremal Regions* (MSERs) [2] were used as distinguished regions. All affine invariant constructions of local frames described in Section 3 were applied. Due to lack of space, matching results are presented for only two matching problems shown in Figures 2 and 3. Table 1 summarizes the number of detected

Figure 3. Valbonne image pair.

distinguished regions, the total number of local frames, the number of established tentative correspondences, and the number of correspondences that were found to be consistent with the epipolar geometry.

Turning our attention to Table 1, we see that the ratio of the number of affine frames and the number of distinguished regions is between 3 and 4, so a highly redundant representation is obtained. The direct comparison of normalised image measurements is a very selective process of tentative correspondences formation. Only approximately fifty (Valbonne) and hundred (Bookshelf) corespondences are selected from the thousands of potential correspondences. However, their number is sufficient for epipolar geometry computation, and the relatively high proportion of EG consistent correspondences (circa 40%) guarantees fast RANSAC termination. In comparison, using a matching procedure based on generalised colour moments, only circa 16% of tentative correspondences were found EG consistent for the Valbonne pair (starting with identical DRs) [2]. Experiments suggest that it is always appropriate to base the correspondences on comparison of full image function, and view any invariant matching only as a preliminary test.

Figure 3 and the middle row of Figure 2 show the epipolar geometry superimposed. The quality of the geometry can be appreciated looking at the close-up in the bottom row of Figure 2. The difference in image resolution is clear. The first row of Figure 2 shows the detected distinguished regions, presenting the elements of the images that are put into correspondence.

6. Conclusions

In this paper, a novel procedure for establishing wide-baseline correspondence was introduced. Starting from robustly detected distinguished regions of data-dependent shape, several processes for local affine frame detection were proposed, proved affine covariant, and experimentally shown to be stable.

The potential of the proposed approach was demonstrated on two wide-baseline matching problems. In the first experiment, correspondence between two views taken from different viewpoints and at very different scales was reliably established. Under such scale change (approximately 3.3 times), the high-resolution image covers less than 10% of the low resolution one. The high number of regions consistent with the estimated epipolar geometry, and the high fraction of inliers in the set of tentative correspondences suggest the limits of the approach have not been reached.

References

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

[2] J. Matas, O. Chum, M. Urban, and T. 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.

[3] C. Mikolajczyk and C. Schmid. Indexing based on scale invariant points. In *ICCV*, 2001.

[4] J. L. Mundy and A. Zisserman, editors. *Geometric Invariance in Computer Vision*. The MIT Press, 1992.

[5] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV'98*, pages 754–760, 1998.

[6] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Proc. 8th International Conference on Computer Vision, Vancouver, Canada*, July 2001.

[7] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[8] T. Suk and J. Flusser. Convex layers: A new tool for recognition of projectively deformed point sets. In F. Solina and A. 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.

[9] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV (1)*, pages 814–828, 2000.

[10] T. Tuytelaars and L. J. V. Gool. Content-based image retrieval based on local affinely invariant regions. In *Visual Information and Information Systems*, pages 493–500, 1999.

[11] Z. Zhang, R. Deriche, O. D. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.