BMVC 2002

Robust Wide Baseline Stereo from Maximally Stable Extremal Regions

J. Matas[1,2], O. Chum[1], M. Urban[1], T. Pajdla[1]

[1]Center for Machine Perception, Dept. of Cybernetics, CTU Prague, Karlovo nám 13, CZ 121 35

[2]CVSSP, University of Surrey, Guildford GU2 7XH, UK

[matas, chum]@cmp.felk.cvut.cz

Abstract

The wide-baseline stereo problem, i.e. the problem of establishing correspondences between a pair of images taken from different viewpoints is studied.

A new set of image elements that are put into correspondence, the so called *extremal regions*, is introduced. Extremal regions possess highly desirable properties: the set is closed under 1. continuous (and thus projective) transformation of image coordinates and 2. monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm (near frame rate) is presented for an affinely-invariant stable subset of extremal regions, the maximally stable extremal regions (MSER).

A new robust similarity measure for establishing tentative correspondences is proposed. The robustness ensures that invariants from multiple measurement regions (regions obtained by invariant constructions from extremal regions), some that are significantly larger (and hence discriminative) than the MSERs, may be used to establish tentative correspondences.

The high utility of MSERs, multiple measurement regions and the robust metric is demonstrated in wide-baseline experiments on image pairs from both indoor and outdoor scenes. Significant change of scale ($3.5\times$), illumination conditions, out-of-plane rotation, occlusion , locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Good estimates of epipolar geometry (average distance from corresponding points to the epipolar line below 0.09 of the inter-pixel distance) are obtained.

1 Introduction

Finding reliable correspondences in two images of a scene taken from arbitrary viewpoints viewed with possibly different cameras and in different illumination conditions is a difficult and critical step towards fully automatic reconstruction of 3D scenes [5]. A crucial issue is *the choice of elements whose correspondence is sought*. In the wide-baseline set-up, local image deformations cannot be realistically approximated by translation or translation with rotation and a full affine model is required. Correspondence cannot be therefore established by comparing regions of a fixed (Euclidean) shape like rectangles or circles since their shape is not preserved under affine transformation.

In most images there are regions that can be detected with high repeatability since they posses some distinguishing, invariant and stable property. We argue that such regions of,

in general, data-dependent shape, called *distinguished regions* (DRs) in the paper, may serve as the elements to be put into correspondence either in stereo matching or object recognition.

The first contribution of the paper is the introduction of a new set of distinguished regions, the so called *extremal regions*. Extremal regions have two desirable properties. The set is closed under continuous (and thus perspective) transformation of image coordinates and, secondly, it is closed under monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm is presented for an affinely-invariant stable subset of extremal regions, the maximally stable extremal regions (MSER). Robustness of a particular type of DR depends on the image data and must be tested experimentally. Successful wide-baseline experiments on indoor and outdoor datasets presented in Section 4 demonstrate the potential of MSERs.

Reliable extraction of a manageable number of potentially corresponding image elements is a necessary but certainly not a sufficient prerequisite for successful wide-baseline matching. With two sets of distinguished regions, the matching problem can be posed as a search in the correspondence space [3]. Forming a complete bipartite graph on the two sets of DRs and searching for a globally consistent subset of correspondences is clearly out of question for computational reasons. Recently, a whole class of stereo matching and object recognition algorithms with common structure has emerged [9, 15, 1, 16, 2, 13, 7, 6]. These methods exploit *local invariant descriptors* to limit the number of tentative correspondences. Important design decisions at this stage include: 1. the choice of measurement regions, i.e. the parts of the image on which invariants are computed, 2. the method of selecting tentative correspondences given the invariant description and 3. the choice of invariants.

Typically, distinguished regions or their scaled version serve as measurement regions and tentative correspondences are established by comparing invariants using Mahalanobis distance [10, 16, 11]. As a second novelty of the presented approach, a robust similarity measure for establishing tentative correspondences is proposed to replace the Mahalanobis distance. The robustness of the proposed similarity measure allows us to use invariants from a collection of measurement regions, even some that are much larger than the associated distinguished region. Measurements from large regions are either very discriminative (it is very unlikely that two large parts of the image are identical) or completely wrong (e.g. if orientation or depth discontinuity becomes part of the region). The former helps establishing reliable tentative (local) correspondences, the influence of the latter is limited due to the robustness of the approach.

Finding epipolar geometry consistent with the largest number of tentative (local) correspondences is the final step of all wide-baseline algorithms. RANSAC has been by far the most widely adopted method since [14]. The presented algorithm takes novel steps to increase the number of matched regions and the precision of the epipolar geometry. The rough epipolar geometry estimated from tentative correspondences is used to guide the search for further region matches. It restricts location to epipolar lines and provides an estimate of affine mapping between corresponding regions. This mapping allows the use of correlation to filter out mismatches. The process significantly increases precision of the EG estimate; the final average inlier distance-from-epipolar-line is below 0.1 pixel. For details see Section 3.

Related work. Since the influential paper by Schmid and Mohr [11] many image matching and wide-baseline stereo algorithms have been proposed, most commonly using

Image I is a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Extremal regions are well defined on images if:

1. $\mathcal{S}$ is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. In this paper only $S = \{0, 1, \ldots, 255\}$ is considered, but extremal regions can be defined on e.g. real-valued images ($\mathcal{S} = R$).
2. An adjacency (neighbourhood) relation $A \subset \mathcal{D} \times \mathcal{D}$ is defined. In this paper 4-neighbourhoods are used, i.e. $p, q \in \mathcal{D}$ are adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$.

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$.

(Outer) Region Boundary $\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$, i.e. the boundary $\partial\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

Extremal Region $\mathcal{Q} \subset D$ is a region such that for all $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Maximally Stable Extremal Region (MSER). Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \ldots$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i^*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}|/|\mathcal{Q}_i|$ has a local minimum at i^* ($|.|$ denotes cardinality). $\Delta \in \mathcal{S}$ is a parameter of the method.

Table 1: **Definitions** used in Section 2

Harris interest points as distinguished regions. Tell and Carlsson [13] proposed a method where line segments connecting Harris interest points form measurement regions. The measurements are characterised by scale invariant Fourier coefficients. The Harris interest detector is stable over a range of scales, but defines no scale or affine invariant measurement region. Baumberg [1] applied an iterative scheme originally proposed by Lindeberg and Garding to associate affine-invariant measurement regions with Harris interest points. In [7], Mikolajczyk and Schmid show that a scale-invariant MR can be found around Harris interest points. In [9], Pritchett and Zisserman form groups of line segments and estimate local homographies using parallelograms as measurement regions. Tuytelaars and Van Gool introduced two new classes of affine-invariant distinguished regions, one based on local intensity extrema [16] the other using point and curve features [15]. In the latter approach, DRs are characterised by measurements from inside an ellipse, constructed in an affine invariant manner. Lowe [6] describes the 'Scale Invariant Feature Transform' approach which produces a scale and orientation-invariant characterisation of interest points.

The rest of the paper is structured as follows. Maximally Stable Extremal Regions are defined and their detection algorithm is described in Section 2. In Section 3, details of a novel robust matching algorithm are given. Experimental results on outdoor and indoor images taken with an uncalibrated camera are presented in Section 4. Presented experiments are summarized and the contributions of the paper are reviewed in Section 5.

2 Maximally Stable Extremal Regions

In this section, we introduce a new type of image elements useful in wide-baseline matching — the *Maximally Stable Extremal Regions*. The regions are defined solely by an extremal property of the intensity function in the region and on its outer boundary.

The concept can be explained informally as follows. Imagine all possible thresholdings of a gray-level image I. We will refer to the pixels below a threshold as 'black' and

to those above or equal as 'white'. If we were shown a movie of thresholded images I_t, with frame t corresponding to threshold t, we would see first a white image. Subsequently black spots corresponding to local intensity minima will appear and grow. At some point regions corresponding to two local minima will merge. Finally, the last image will be black. The set of all connected components of all frames of the movie is the set of all maximal regions; minimal regions could be obtained by inverting the intensity of I and running the same process. The formal definition of the MSER concept and the necessary auxiliary definitions are given in Table 1.

In many images, local binarization is stable over a large range of thresholds in certain regions. Such regions are of interest since they posses the following properties:

- **Invariance to affine transformation of image intensities.**
- **Covariance to adjacency preserving** (continuous) transformation $T : \mathcal{D} \to \mathcal{D}$ on the image domain.
- **Stability**, since only extremal regions whose support is virtually unchanged over a range of thresholds is selected.
- **Multi-scale detection.** Since no smoothing is involved, both very fine and very large structure is detected.
- The set of all extremal regions can be **enumerated in** $O(n \log\log n)$, where n is the number of pixels in the image.

Enumeration of extremal regions proceeds as follows. First, pixels are sorted by intensity. The computational complexity of this step is $\mathcal{O}(n)$ if the range of image values $\mathcal{S}$ is small, e.g. the typical $\{0, \ldots, 255\}$, since the sort can be implemented as BINSORT [12]. After sorting, pixels are placed in the image (either in decreasing or increasing order) and the list of connected components and their areas is maintained using the efficient union-find algorithm [12]. The complexity of our union-find implementation is $\mathcal{O}(n \log\log n)$, i.e. almost linear[1]. Importantly, the algorithm is very fast in practice. The MSER detection takes only 0.14 seconds on a Linux PC with the Athlon XP 1600+ processor for an 530x350 image ($n = 185500$).

The process produces a data structure storing the area of each connected component as a function of intensity. A merge of two components is viewed as termination of existence of the smaller component and an insertion of all pixels of the smaller component into the larger one. Finally, intensity levels that are local minima of the rate of change of the area function are selected as thresholds producing maximally stable extremal regions. In the output, each MSER is represented by position of a local intensity minimum (or maximum) and a threshold.

Notes. The structure of the above algorithm and of an efficient **watershed algorithm** [17] is essentially identical. However, the structure of the *output* of the two algorithms is different. The watershed is a partitioning of $\mathcal{D}$, i.e. a set of regions $\mathcal{R}_i : \bigcup \mathcal{R}_i = D, \mathcal{R}_j \cap \mathcal{R}_k = \emptyset$. In watershed computation, focus is on the thresholds where regions merge (and two watersheds touch). Such threshold are of little interest here, since they are highly unstable – after merge, the region area jumps. In MSER detection, we seek a range of thresholds that leaves the watershed basin effectively unchanged. Detection of MSER is also related to **thresholding**. Every extremal region is a connected component of a

[1] even faster (but more complex) connected component algorithms exist with $O(n\alpha(n))$ complexity, where α is the inverse Ackerman function; $\alpha(n) \leq 4$ for all practical n.

thresholded image. However, no global or 'optimal' threshold is sought, all thresholds are tested and the stability of the connected components evaluated. The output of the MSER detector is not a binarized image. For some parts of the image, multiple stable thresholds exist and a system of nested subsets is output in this case. Finally we remark that MSERs can be defined on any image (even high-dimensional) whose pixel values are from a totally ordered set.

3 The proposed robust wide-baseline algorithm

Distinguished region detection. As a first step, the DRs are detected - the MSERs computed on the intensity image (MSER+) and on the inverted image (MSER-).

Measurement regions. A measurement region of arbitrary size may be associated with each DR, if the construction is affine-covariant. Smaller measurement regions are both more likely to satisfy the planarity condition and not to cross a discontinuity in depth or orientation. On the other hand, small regions are less discriminative, i. e. they are much less likely to be unique. Increasing the size of a measurement region carries the risk of including parts of background that are completely different in the two images considered. Clearly, the optimal size of a MR depends on the scene content and it is different for each DR. In [16], Tuytelaars at al. double the elliptical DR to increase discriminability, while keeping the probability of crossing object boundaries at an acceptable level.

In the proposed algorithm, measurement regions are selected at multiple scales: the DR itself, 1.5, 2 and 3 times scaled convex hull of the DR. Since matching is accomplished in a robust manner, we benefit from the increase of distinctiveness of large regions without being severely affected by clutter or non-planarity of the DR's pre-image. This is a novelty of our approach. Commonly, Mahalanobis distance has been used in MR matching. However, the non-robustness of this metric means that matching may fail because of a single corrupted measurement (this happened in the experiments reported below).

Invariant description. In all experiments, rotational invariants (based on complex moments) were used after applying a transformation that diagonalises the regions covariance matrix of the DR. In combination, this is an affinely-invariant procedure. Combination of rotational and affinely invariant generalised colour moments [8] gave a similar result. On their own, these affine invariants failed on problems with a large scale change.

Robust matching. A measurement taken from an almost planar patch of the scene with stable invariant description will be referred to as a 'good measurement'. Unstable measurements or those computed on non-planar surfaces or at discontinuities in depth or orientation will be referred to as 'corrupted measurements'.

The robust similarity is computed as follows. For each measurement M_A^i on region A, k regions $B_1, \ldots, B_k$ from the other image with the corresponding i-th measurement $M_{B_1}^i, \ldots, M_{B_k}^i$ nearest to M_A^i are found and a vote is cast suggesting correspondence of A and each of $B_1, \ldots, B_k$.

Votes are summed over all measurements. In the current implementation 216 invariants at each scale, i.e. a total of 864 measurements are used ($i \in [1, 864]$). The DRs with the largest number of votes are the candidates for tentative correspondences. Experimentally, we found that k set to 1% of the number of regions gives good results.

Probabilistic analysis of the likelihood of the success of the procedure is not simple, since the distribution of invariants and their noise is image-dependent. We therefore only suppose that corrupted measurements spread their votes randomly, not conspiring to create a high score and that good measurements are more likely to vote for correct matches.

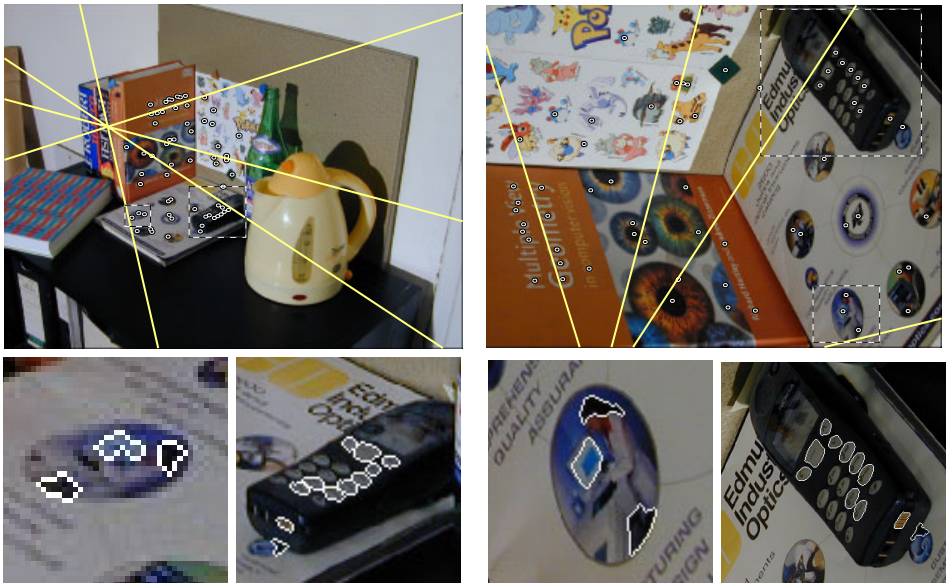

Figure 1: BOOKSHELF: Estimated epipolar geometry on indoor scene with significant scale change. In the cutouts the change in the resolution of detected DRs is clearly visible.

Tentative correspondences using correlation. Invariant description is used as a preliminary test. The final selection of tentative correspondences is based on correlation. First transformations that diagonalise the covariance matrix of the DRs are applied. The resulting circular regions are correlated (for all relative rotations). This procedure is done efficiently in polar coordinates for different sizes of circles.

Rough **epipolar geometry** (EG) is estimated by applying RANSAC to the centers of gravity of DRs. Subsequently, the precision of the EG estimate is significantly improved by the following process. First, an affine transformation between pairs of potentially corresponding DRs, i.e. the DRs consistent with the rough EG, is computed. Correspondence of covariance matrices defines an affine transformation up to a rotation. The rotation is determined from epipolar lines. Next, DR correspondences are pruned and only those with correlation of their transformed images above a threshold are selected. In the next step, RANSAC is applied again, but this time with a very narrow threshold. The final improvement of the EG is achieved by adding to RANSAC inliers DR pairs whose convex hull centres are EG-consistent. Commonly, DRs differ in minute differences that render their centres of gravity inconsistent with the fine EG, but the centers of the convex hulls are precise enough. The precision of the final EG, estimated linearly by the eight point algorithm (without bundle adjustment or radial distortion correction) is surprisingly high. The average distance of inliers from epipolar line is below 0.1 pixel, see Table 3.

4 Experiments

The following experiments were conducted:

Bookshelf, (Fig. 1). The BOOKSHELF scene tests performance under a very large scale change. The corresponding DRs in the left view are confined only to a small part of the

Figure 2: VALBONNE: Estimated epipolar geometry and points associated to the matched regions are shown in the first row. Cutouts in the second row show matched bricks.

image since the rest of the scene is not visible in the second view. Different resolution of detected features is evident in the close-up.
Valbonne, (Fig. 2). This outdoor scene has been analysed in the literature [10, 9]. Repetitive patterns such as bricks are present. The part of the scene visible in both views covers a small fraction of the image.

number of:	MSER -	MSER +	TC
Bookshelf	511 × 908	349 × 488	85
Valbonne	906 × 1012	761 × 950	49
Wash	1026 × 714	542 × 448	171
Kampa	1015 × 914	659 × 652	303
Cyl. Box	1043 × 627	788 × 39	63
Shout	298 × 348	80 × 93	151

Table 2: Number of DRs detected in images. The number of tentative correspondences is given in the TC column.

Wash, (Fig. 3). Results on this image set have been presented in [16]. The camera undergoes significant translation and rotation. The ordering constraint is notably violated, objects appear on different backgrounds.
Kampa, (Fig. 4), is an example of an urban outdoor scene. A relatively large fraction of the images is covered by changing sky. Repeating windows made matching difficult.
Cylindrical Box, (Fig. 5, top and bottom left), shows a metal box on a textured floor. The regions matched on the box demonstrate performance on a non-planar surface. A significant change of illumination and a strong specular reflection is present in the second image that was taken with a flash (this strongly decreases the number of MSER +).
Shout, (Fig. 5, bottom right). This scene has been used in [16]. Since the spectral power distribution of the illumination and the position of light sources is significantly different, we included the test to demonstrate performance in variable illumination conditions.

Results are summarized in Tables 2 and 3. Table 2 shows the number of detected DRs in the left × right images for both types of the DRs (MSER- and MSER+). The number of tentative correspondences is given in the last column of Table 2. Table 3 shows the

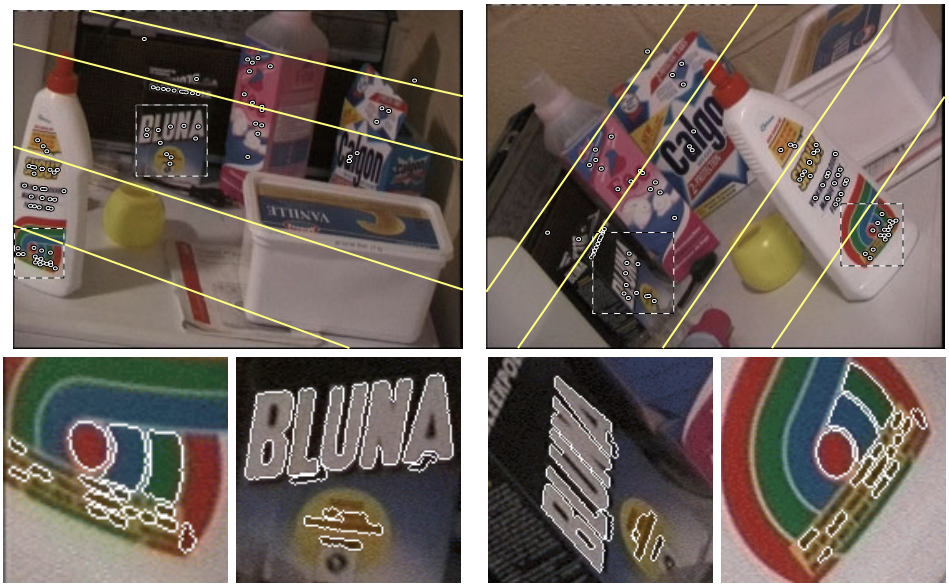

Figure 3: WASH: Epipolar geometry and dense matched regions with fully affine distortion.

	TC	rough EG	rough $d_\perp$	EG + corr	fine EG	fine $d_\perp$	miss
Bookshelf	85	25	0.48	151	63	0.09	1
Valbonne	49	27	0.17	180	82	0.08	0
Wash	171	42	0.34	220	86	0.08	2
Kampa	303	78	0.34	422	185	0.08	2
Cyl. Box	63	23	0.15	102	67	0.09	3
Shout	151	44	0.43	220	86	0.08	1

Table 3: Experimental results. For details see the text, at the beginning of Section 4.

number of correspondences established in different stages of the algorithm. Column 'TC' repeats the number of tentative correspondences. Column 'rough EG' displays the number of tentative correspondences consistent with the rough estimate of the epipolar geometry. The ratio of 'TC' and 'rough EG' determines the speed of the RANSAC algorithm. The column headed 'EG + corr' gives the number of correspondences consistent with rough EG that passed the correlation test. Notice that the numbers are much higher than those in the 'rough EG' column. The final number of correspondences is given in the penultimate column 'fine EG'. Average distances from epipolar lines are presented in columns 'rough $d_\perp$' and 'fine $d_\perp$'. We can see, that the precision of the estimated epipolar geometry is very high, much higher than the precision of the rough EG. The last column shows the number of mismatches (found manually).

5 Conclusions

In the paper, a new method for wide-baseline matching was proposed. The three main novelties are: the introduction of MSERs, robust matching of local features and the use of multiple scaled measurement regions.

Figure 4: Estimated EG on an outdoor scene.

The MSERs are sets of image elements, closed under the affine transformation of image coordinates and invariant to affine transformation of intensity. An efficient (near linear complexity) and practically fast detection algorithm was presented. The stability and high utility of MSERs was demonstrated experimentally. Another novelty of the approach is the use of a *robust similarity measure* for establishing tentative correspondences. Due to the robustness, we were able to consider invariants from *multiple measurement regions*, even some that were significantly larger (and hence probably discriminative) than the associated MSER.

Good estimates of epipolar geometry were obtained on challenging wide-baseline problems with the robustified matching algorithm operating on the output produced by the MSER detector. The average distance from corresponding points to the epipolar line was below 0.09 of the inter-pixel distance. Significant change of scale ($3.5\times$), illumination conditions, out-of-plane rotation, occlusion , locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Test images included both outdoor and indoor scenes, some already used in published work.

In future work, we intend to proceed towards fully automatic projective reconstruction of the 3D scene, which requires computing projective reconstruction and dense matching. Secondly, we will investigate properties of robust similarity measures and their selection based on statistical properties of the data.

References

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

[2] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *CVPR00*, pages I:612–618, 2000.

[3] W. Eric L. Grimson. *Object Recognition*. MIT Press, 1990.

[4] R. Hartley. In defence of the 8-point algorithm. In *ICCV95*, pages 1064–1070, 1995.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.

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

[7] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Eighth Int. Conference on Computer Vision (Vancouver, Canada)*, 2001.

Figure 5: CYLINDRICAL BOX: Epipolar geometry (top) and matched regions (bottom left). Fully affine distortion, a non-planar object, textured surface and a strong specular reflection are present in the scene. SHOUT (bottom right), a scene with a change of illumination spectral power distribution.

[8] F. Mindru, T. Moons, and L.J. van Gool. Recognizing color patterns irrespective of viewpoint and illumination. In *CVPR99*, pages I:368–373, 1999.

[9] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 754–760, January 1998.

[10] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Eighth Int. Conference on Computer Vision (Vancouver, Canada)*, 2001.

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

[12] R. Sedgewick. *Algorithms*. Addison-Wesley, 2nd edition, 1988.

[13] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV00*, 2000.

[14] P.H.S. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. In *BMVC96*, page Motion and Active Vision, 1996.

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

[16] T. Tuytelaars and L. Van Gool. Wide baseline stereo based on local, affinely invariant regions. In M. Mirmehdi and B. Thomas, editors, *Proc British Machine Vision Conference BMVC2000*, pages 412–422, London, UK, 2000.

[17] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, June 1991.

Acknowledgments. The authors were supported by the European Union under project IST-2001-32184 and by the Grant Agency of the Czech Republic under projects GACR 102/01/0971 and GACR 102/02/1539. The SHOUT and WASH images were kindly made available by Tinne Tuytelaars.