Wide-baseline Stereo from Distinguished Regions

Abstract

The problem of establishing correspondences between a pair of images taken from different viewpoints, i.e. the "wide-baseline stereo" problem, is fundamental in computer vision.

In the paper two new wide-baseline stereo algorithms are proposed. The first establishes correspondence of closed loops of Deriche edges. The second is based on junctions of line segments, detected by Hough Transform and a probabilistic relaxation process. Tentative correspondences are selected using invariants computed on regions defined in a quasi-invariant manner. A correct epipolar geometry of image pairs acquired by an uncalibrated camera undergoing a large rotation, change of scale, viewpoint change, and/or occlusion is established completely automatically. The algorithms are tested experimentally on indoor and outdoor real-world scenes.

In a second contribution, we describe and analyse the structure of a class of stereo matching and object recognition algorithms that are based on invariant descriptors computed over regions possessing some distinguishing property. The common elements of the methods are discussed, and the central concepts of distinguished region *and* measurement region *are defined.*

1 Introduction

Finding reliable correspondences in two images of a scene or an object taken from arbitrary viewpoints with possibly different cameras in different illumination conditions is a difficult and critical step on the way towards fully automatic reconstruction of 3D scenes from intensity images. In this paper two new algorithms addressing this type of correspondence problem, often referred to as "wide-baseline stereo", are introduced. The presented methods are experimentally shown to provide good estimates of epipolar geometry for pairs of disparate views of real-world scenes[1] with significant change of scale, camera rotation, and 3D translation of the viewpoint.

In a second contribution, we describe and analyse the structure of a class of stereo matching and object recognition algorithms that are based on local invariant descriptors computed over regions possessing some distinguishing property. The common elements of the methods are discussed, and the central concepts of *distinguished region* and *measurement region* are defined. We postpone review of the related literature until after introducing these concepts since most published wide-baseline matching algorithms and a number of closely related image retrieval and recognition methods such as [12, 17, 1, 18, 2, 15, 8, 7] can be seen as instances of the general framework. These methods whose structure is summarised in Algorithm 1 differ in the particular choice of distinguished and measurement regions.

The generalisation, i.e. model of structure of a number of approaches, makes analysis and comparison of the algorithms easier, highlighting what particular choices were made at what point and what is the set of possible options at each stage. Moreover, many future lines of research become immediately obvious. Secondly, we clearly see convergence of problems addressed in stereo matching and in image retrieval and object recognition, which could lead to mutually beneficial exchange of techniques and perhaps unification of terminology and problem formulation.

The rest of the paper is structured as follows. We start by describing the structure of the matching algorithm based on distinguished regions (Section 2). Next, previous work in the area is reviewed. In Section 4, two new wide-baseline stereo algorithms are presented. The first establishes correspondence of closed loops of Deriche edges. The second is based on junctions of line segments, detected by Hough Transform and a probabilistic relaxation process. In both methods, the RANSAC procedure [4, 16] is employed to find a large consistent set of matches. Experimental results on outdoor and indoor images taken with an uncalibrated camera are presented in Section 5. The paper is concluded in Section 6.

2 Correspondence from Distinguished Regions

Imagine you are presented with two images, depicting at least partially the same scene (object) taken with possibly different cameras from two arbitrarily different viewpoints. You are asked to mark corresponding points in the image pair. We would argue that, unless *distinguished regions* are present in the two images, the task is extremely hard. Two

[1] The generalisation of the proposed method to multiple views is straightforward.

Figure 1: Examples of *distinguished regions*, on the ANACIN box and the CAT from the COIL-20 database [11]

views of a white featureless wall, a patch of grass, sea surface or an ant hill might be good examples. On the other hand, look at objects from the COIL-20 database depicted in Figure 1. On most object, we find surface patches that can be separated from their surroundings. The eye in the CAT image or the letters on the ANACIN box are two such regions. The different views of the CAT and ANACIN images suggest that the regions marked in Figure 1 are detectable (e.g. by a watershed-like process or local thresholding) over a wide range of views. Before proceeding further, we give a more formal definition:

Definition 1 A Distinguished Region *(DR) is any subset of an image that is a projection of a part of a scene possessing a distinguishing property allowing its detection (segmentation, figure-ground separation) over a range of viewpoints and illumination conditions.*

In other words, DR detection must be repeatable and stable w.r.t. viewpoint and illumination change. Concrete examples of DR detectors used in wide-base line stereo will be discussed later. Note that we do not require DRs to have some transformation-invariant property that is unique in the image. If a DR possesses such a property, finding its corresponding DR (if it is visible in the other image) is greatly simplified. To increase the likelihood of this desirable situation, DRs can be equipped with a characterisation computed on associated *measurement regions*.

Definition 2 A Measurement Region *(MR) is any subset of an image defined by a transformation-invariant construction (projective, affine, similarity invariant) from one or more (in case of grouping) distinguished regions.*

Since DRs are projections of the same part of the scene in both views and MRs are defined in a transformation-invariant manner they are quasi viewpoint-invariant. Besides the simplest and most common case where the MR is the DR itself, a MR may be constructed for example as: a convex hull of a DR (perspectively invariant, see Algorithm 2), a fitted ellipse (affinely invariant, [17]), a line segment between a pair of interest points [15] or any region defined in DR-derived coordinates. Of course, invariant measurement from a single or even multiple MRs associated with a DR will not guarantee a unique match on e.g. repetitive patterns. But often, as in the case of above-mentioned DRs in Figure 1, DR characterisation by invariants computed on MRs might be unique or almost unique. The neighbourhood of the marked region around letter A on the ANACIN box is unique. The characterisation of the eye region on the CAT image might be similar to a few, but not many, neighbourhoods around other circular black-on-yellow spots.

Characterisation. The most simple situation arises if a local affine frame is defined on the DR. Then a MR could be any region specified in terms of the local coordinate frame. Photometrically normalised pixel values from a normalised patch characterise the DR invariantly. More commonly, only a point or a point and a scale factor are known, and rotation invariants [14, 13] or affine invariants must be used [18].

Tentative Correspondences. At this stage, we have a set of DRs for each image and a potentially large number of invariant measurements associated with each DR. For problems of realistic size it is not practical to assume that any pair of DRs may be in correspondence and verify the hypothesis by checking geometric consistency, since an astronomical number of hypotheses would have to be considered. Instead, only correspondences with similar characterisation are selected. Selecting mutually nearest pairs in Mahalanobis distance is the most common method [13, 18, 14]. We adopted the minimal Mahalanobis distance method in the algorithms presented later (despite the fact that the assumptions under which the method would be optimal are almost certainly not met [5]). Note that the objective of this stage is not to keep the maximum possible number of good correspondences, but rather to maximise the fraction of good correspondences. The fraction determines the speed of epipolar geometry estimation by the RANSAC procedure [16].

Epipolar Geometry estimation is carried out by a robust statistical method, most commonly RANSAC. In RANSAC, randomly selected subsets of tentative correspondences instantiate an epipolar geometry model. The number of correspondences consistent with the model defines its quality. The hypothesise - verify loop is terminated when the likelihood of finding a better model falls below a predefined threshold. The process of establishing correspondences from distinguished regions is summarised in Algorithm 1:

Algorithm 1: Wide-baseline Stereo from Distinguished Regions - The Framework

1. Detect *Distinguished Regions*.
2. Detect *Measurement Regions*.
3. Characterise DRs via invariant measurements on MRs.
4. Establish Tentative Correspondences.
5. Estimate Epipolar Geometry.

Having outlined the framework, we make remarks on properties of DRs and MRs:

- Terminology. The *distinguished regions* are referred to in the literature as 'interest points' [2], 'features' [1] or 'invariant regions' [18] (in [18], 'invariant features' refer to the invariant characterisation). After considerations we rejected the term 'invariant region', since (ignoring visibility issues) every locally planar image patch has a corresponding patch in the second image with the same pre-image. But for a vast majority of regions, it is not distinguishable.

- We adopted the terms distinguished and measurement *regions*. However any set of pixels, not necessarily contiguous, can posses a distinguishing property. Many perceptual grouping processes detect such arrangements. e.g. a set of (unconnected) edgels lying along a straight line form a DR of maximum of edgel density. The property is view-point quasi-invariant and detectable by the Hough Transform. The 'distinguished pixel set' would be a more precise term, but it is cumbersome.

- The separation of the concepts of DR and MRs is important and not made explicit in the literature. For instance, Tuytelaars and Van Gool wrote [18]: "'Invariant regions' are image patches that automatically deform with changing viewpoint as to keep on covering identical physical part of a scene. Such regions are then described by a set of invariant features ".

- Clearly, different DR detectors (based on colour, texture, edges, local shape, etc.) are suitable for different images. Multiple DR types, multiple MR constructions and characterisations may be used in steps 1 to 3 of Algorithm 1. In general, since there is no way of knowing what type of DRs is suitable for an a priori unknown scene, it seems natural to adopt an opportunistic approach and to try as many as possible.

- We are not suggesting to "segment the image", but perhaps to segment (select, separate) the segmentable (i.e. the distinguishable). The "interpretation" of the distinguished regions is irrelevant; repeatable detection in varying acquisition conditions is important.

- Automatic, statistical approaches and learning techniques may be applied in DR detector design since repeatability tests do not require manually labelled data. The relationship to appearance-based methods [10] is clear. In both cases measurements derived directly from intensities are used but here the problem of *where to measure* is central.

- All geometric constraints used in the process of epipolar geometry estimation come from DRs; the geometry of MR in our framework is completely determined by DR and hence adds no multi-view constraints.

3 Previous Work

Since the influential paper by Schmid and Mohr [14] many image matching and wide-baseline stereo algorithms have used Harris interest points as distinguished regions. Tell and Carlsson [15] proposed a method where line segments connecting Harris interest points form measurement regions. The MRs are characterised by scale invariant Fourier coefficients. Harris interest point detector is only rotationally invariant. Its maximum response 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 [8], Mikolajczyk and Schmid show that a scale-invariant MR can be found around Harris interest points.

In [12], 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 distinguished regions, one based on a point and curve [17], the second on local intensity extrema [18]. Both methods are affine invariant. Lowe [7] describes the 'Scale Invariant Feature Transform' approach which produces a scale and orientation-invariant characterisation of interest points.

4 Two new wide-baseline algorithms

In this section, two new instances of the above-mentioned general approach are presented. The implementation is kept very simple, to show clearly the concepts introduced - there is space for improvement in every step of the algorithm.

Two types of distinguished regions are exploited - closed loops of intensity edges and junctions of line segments detected by Hough Transform from intensity edges. Edges are computed using Deriche edge detector [3].

In the case of closed loops, the measurement region is the set of pixels inside the convex hull of a closed loop (convex hull is preserved under oriented projective transform [6]). RGB values in the measurement region are characterised by 21 generalised colour moment invariants [9] of degree 0 and 1 in spatial coordinates and 0, 1, 2 in RGB coordinates. This description is invariant to affine transformation of pixel coordinates and to scaling of the brightness function in each colour band.

The initial set of tentative matches is formed as follows. Two regions l and r are taken as a candidate for a match iff region l from the "left" image L is the closest region to r in the "right" image R and *vice-versa*. The distance between regions is measured by Mahalanobis distance in 21-dimensional feature space. We call such a pair *mutually nearest*.

In the case of junctions, the measurement region is specified in terms of a local affine coordinate frame defined by three points of the junction. The initial set of tentative matches is formed as described above. The distance is computed as cross-correlation of RGB values in the normalised coordinate frame.

Finally, a RANSAC algorithm [16] finds a subset of geometrically consistent matches and estimates the best model describing the geometry of the stereo pair (i.e. model of an epipolar geometry, an affine transformation, or a homography). The geometric entities input into RANSAC are the centres of gravity for the closed loops and triplets of points for the junctions. The two methods are summarised in Algorithms 2 and 3. The intermediate data that are produced by Algorithms 2 and 3 are visualised in Figs. 2 and 3 respectively.

Algorithm 2: Algorithm based on closed loops

1. **DR**: closed loops of Deriche edges.
2. **MR**: convex hull of DR.
3. **Characterisation:**
   21 affine invariants based on generalised colour moments computed on a convex hull of DRs.

$$M_{pq}^{abc} = \iint x^p y^q R(x,y)^a G(x,y)^b B(x,y)^c dxdy$$

4. **Tentative Correspondences**: DRs with mutually nearest characterisation (in Mahalanobis distance).
5. **Epipolar Geometry** estimated by RANSAC. Each DR provides a single point-to-point correspondence (centres of gravity of the closed loop).

Algorithm 3: Algorithm based on junctions

1. **DR**: junctions of line segments.
2. **MR**: a parallelogram defined by 3 points on line segments forming the junction.
3. **Characterisation:** normalised image patch derived from the affine frame of a junction.
4. **Tentative Correspondences**: DRs with mutually nearest characterisation (cross-correlation of normalised RGB values).
5. **Epipolar Geometry** estimated by RANSAC. Each DR provides 3 point-to-point constraints.

5 Experiments

The behaviour of Algorithms 2 and 3 was tested on two types of scenes. Experiment A was carried out on four pairs of images of an indoor scene (see Figs. 4, 5, 6, 7). In Experiment B, images of an outdoor scene were used (see Fig. 8). Performance of Algorithms 2 is demonstrated on the indoor scene. Results of Algorithm 3 are presented for the outdoor scene[2].

Experiment A: Indoor images. The viewpoint change in each of the four indoor experiments differed; the transformations are described in Table 1. Images were acquired by a standard digital camera with 640×480 pixel resolution. The algorithm was run with the same parameter settings in all four cases. In each image, more than 100 closed loops were recovered. The exact number of closed loops in the left and right images are shown in Table 2. If every pair of distinguished region were considered, about 25 000 tentative correspondences would be formed. The numbers of such pairs for experiments A-1 to A-4 are shown in the second column of Table 3. Attempting to select a subset with seven correct correspondences by random sampling in the unpruned set is hopeless (the epipolar geometry was estimated by the 7-point algorithm [6]). Tentative correspondences comprised only of those pairs whose general colour invariant [9] were mutually nearest. The number of correspondences of distinguished region pairs is shown in the third column of Table 3.

The expected runtime of RANSAC depends on the number of hypotheses tested, which is $< \epsilon^{-m}$, where ϵ is the probability of selecting a correspondence that is correct and

[2] successful experiments with Alg. 3 on indoor images are not reported in this paper for lack of space.

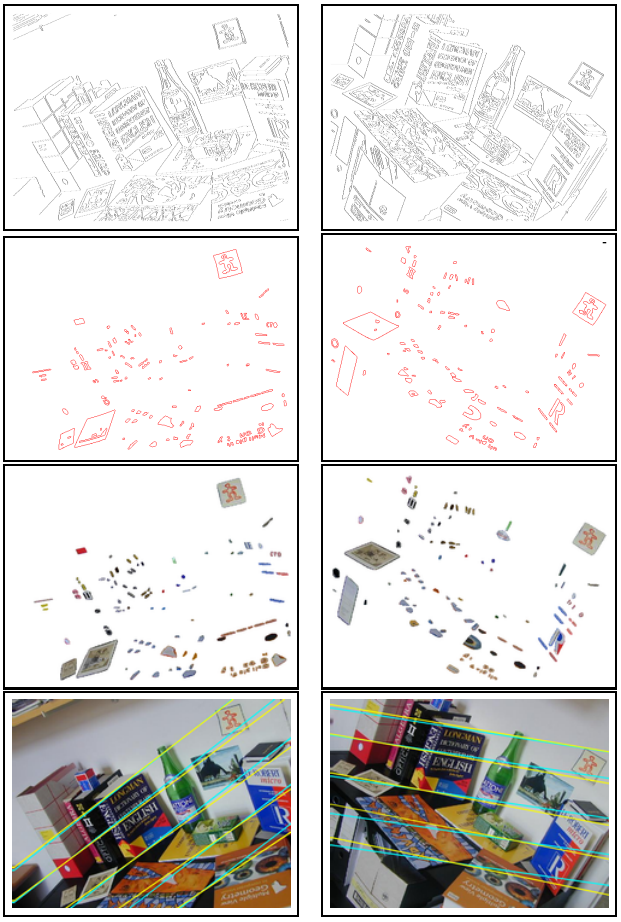

Figure 2: Intermediate results of Algorithm 2. From top to bottom: Deriche edges, closed loops, convex hulls of the closed loops and the original images with estimated epipolar lines.

Exp.	viewpoint transformations
A-1	translation and rotation
A-2	translation and rotation
A-3	0.75 scale change by zooming
A-4	translation, rotation, horizontal flip, occlusion

Table 1: Viewpoint transformations in Experiment A.

m is the number of point-to-point correspondences needed to compute epipolar geometry; in our case $m = 7$. Overall

	image 1	image 2
A-1	156	177
A-2	173	144
A-3	163	156
A-4	132	159

Table 2: Experiment A. Number of detected closed loops.

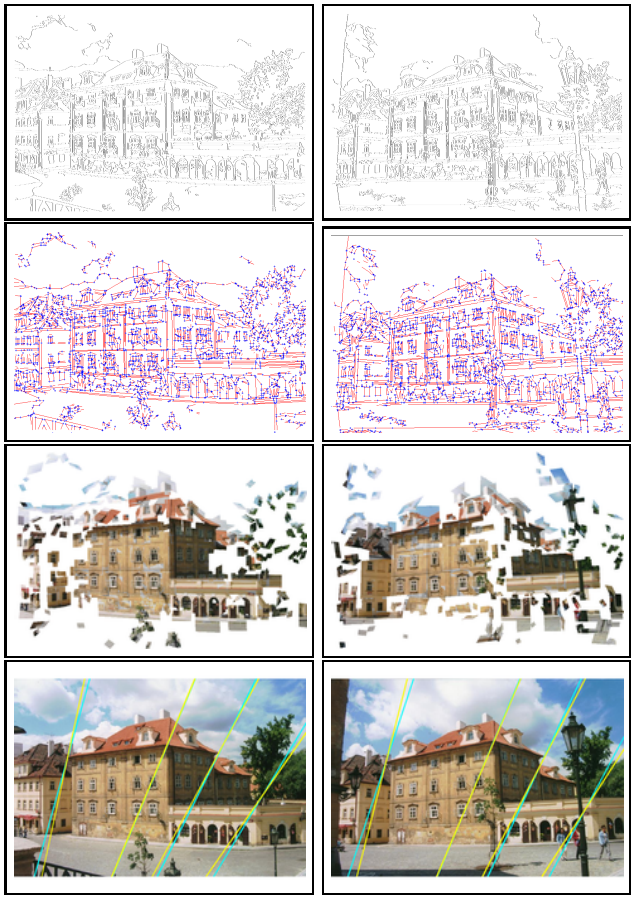

Figure 3: Results of algorithm 3. From top to bottom: Deriche edges, junctions, measurement regions and the original images with estimated epipolar lines.

in experiment A, the number of subset chosen and epipolar geometries tested was on average $\approx$ 300, which is a manageable number. By inspection we found that the percentage of correct matches in the pruned set ranged from 33% to 52%, see the rightmost column of Table 3. For these percentages of correct matches the expected number of RANSAC verifications ranges from $1/0.52^7 \approx 100$ to $1/0.33^7 \approx 2500$, which is consistent with the observed runtime.

To demonstrate that the epipolar geometry estimated from tentative pairs is roughly correct, we have also estimated epipolar geometry from manually selected correspondences. Figures 4, 5, 7 depict epipolar lines of manual and computed epipolar geometries. The epipolar geometry in experiments A-1,A-2, and A-4 is roughly correct. Our objective is not high precision; the epipolar geometry of the wide-baseline algorithm may be improved e.g. by iterative bundle adjustment. The affine transformation of the scale change (experiment A-3) was also correctly recovered (Fig.

6). Since both images of experiment A-3 were taken from the same viewpoint, we did not attempt to recover epipolar geometry.

Experiment B: Outdoor scene. The pair of images used in the experiment was acquired by a still camera and then scanned to a 800×540 pixel resolution. In each image, more than 1000 junctions were found by a probabilistic relaxation algorithm. Around 400 junctions with sufficient quality (the quality is an output of the relaxation process) formed the distinguished regions. The measurement region was defined in terms of a local affine frame. The three points needed to define the frame were the intersection and the other two ends (i.e. those not near the intersection) of line segments forming the junction.

For each junction pair, cross-correlation of the geometrically normalised image patches was computed. 117 tentative matches were found as described in Section 4. The number of subsets chosen (and epipolar geometries tested) was on average ≈ 50. Since each junction provided 3 point-to-point correspondences and therefore only subsets of size 3 had to be chosen, the RANSAC procedure was much faster than in Experiment A. With nine correspondences, the epipolar geometry was obtained by a least square algorithm [6]. By inspection we found that 41, i.e. approximately 35% of junction correspondence were correct (see Table 4) The percentage of correct correspondence predicts RANSAC run-time in the order of $1/0.35^3 \approx 25$, which is consistent with the observed run-time. Figure 8 depicts epipolar lines of epipolar geometries estimated both automatically and manually.

6 Conclusions

In the paper, two new wide-baseline stereo algorithms were proposed. The first establishes correspondence of closed loops of Deriche edges. The second is based on junctions of line segments, detected by Hough Transform and a probabilistic relaxation process. Epipolar geometry of image pairs showing significant rotation and scale variation was successfully recovered. Outdoor and indoor images taken with an uncalibrated camera were used in the experiment.

In a second contribution, we describe and analyse the structure of a class of stereo matching and object recognition algorithms that are based on invariant descriptors computed over regions possessing some distinguishing property. The common elements of the methods are discussed, and the central concepts of *distinguished region* and *measurement region* are defined. We believe that many new wide-baseline methods can be obtained following the framework, each increasing the range of scene types for which correspondence can be established automatically.

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] O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, Massachusetts, 1993.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381, June 1981.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing. Academic Press, London, Great Britain, 2nd edition, 1990.

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

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

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

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

[10] S. K. Nayar and T. Poggio, editors. *Early Visual Learning*. Oxford University Press, 1996.

[11] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). Technical report, Columbia University, 1996.

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

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

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

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

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

[17] 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.

[18] 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.

Exp.	pairs		
	all	selected	% correct
A-1	27612	52	44%
A-2	24912	50	52%
A-3	25428	42	33%
A-4	20988	46	39%

Table 3: Experiment A. Number of correspondences.

DR in image 1	DR in image 2	all pairs	selected pairs	% correct
414	357	147798	117	35

Table 4: Experiment. B. Number of junctions and number of selected pairs.

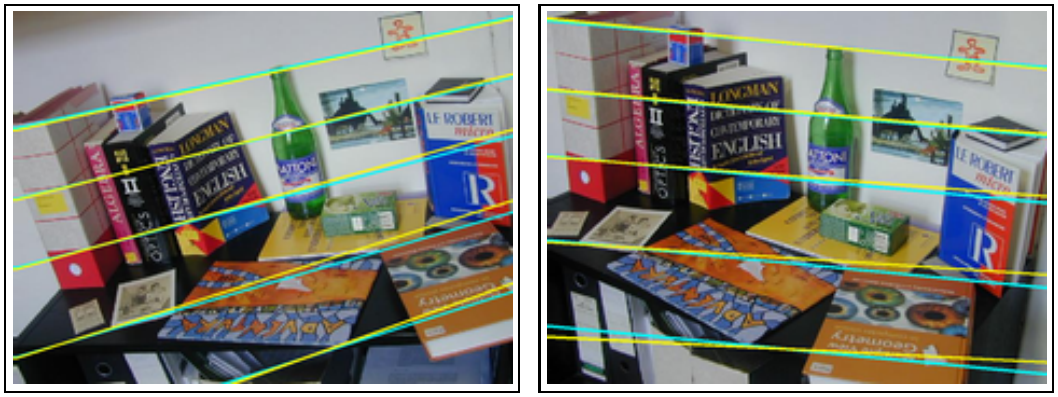

Figure 4: Experiment A-1. Epipolar lines of estimated (yellow) and manual (blue) epipolar geometry.

Figure 5: Experiment A-2. Epipolar lines of estimated (yellow) and manual (blue) epipolar geometry.

Figure 6: Experiment A-3. Recovered inliers of affine transformation.

Figure 7: Experiment A-4. Epipolar lines of estimated (yellow) and manual (blue) epipolar geometry. The black stripe models partial occlusion.

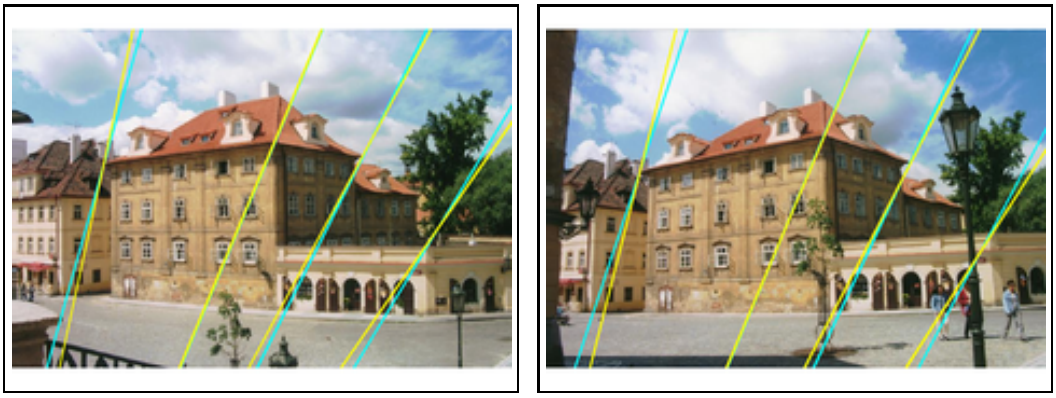

Figure 8: Experiment B-1. Epipolar lines of estimated (yellow) and manual (blue) epipolar geometry.