CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY

RESEARCH REPORT

ISSN 1213-2365

Distinguished Regions for Wide-baseline Stereo

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

[1] Center for Machine Perception,
Department of Cybernetics
Faculty of Electrical Engineering,
Czech Technical University in Prague

166 27 Prague 6, Technicka 2
Czech Republic

[2] Centre for Vision Speech and Signal Proc.
School of Electronics, Computing and
Mathematics
University of Surrey

Guildford, GU2 7XH
United Kingdom

CTU–CMP–2001–33

November 23, 2001

The author was supported by the Czech Ministry of Education under project MSM 210000012 and by the Grant Agency of the Czech Republic under project GACR 102/00/1679.

Research Reports of CMP, Czech Technical University in Prague, No. 33, 2001

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Abstract

The problem of establishing correspondences between a pair of images taken from different viewpoints, i.e. the "wide-baseline stereo" problem, is studied in the paper. The choice of image elements that are put into correspondence in the wide-baseline matching problem is discussed. The concept of a *distinguished region* is introduced and formally defined and it is argued distinguished regions are very good candidates for matching.

Two new types of distinguished regions, the *Separated Elementary Cycles of the Edge Graph (SECs)* and the *Maximally Stable Extremal Regions (MSERs)*, are introduced. For both types, an efficient (near linear complexity) and practically fast detection algorithm is presented. Experimentally the stability of the proposed DRs is shown on disparate views of real-world scenes with significant change of scale, camera rotation and 3D translation of the viewpoint.

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

In experiments on indoor and outdoor image pairs, good estimates of epipolar geometry are obtained on challenging wide-baseline problems with the robustified matching algorithm operating on the output produced by the proposed detectors of distinguished regions. Locally fully affine distortions and significant occlusion were present in the tests.

1 Introduction

Finding reliable correspondences in two images of a scene taken from arbitrary viewpoints with possibly different cameras in different illumination conditions is a difficult and critical step towards fully automatic reconstruction of 3D scenes [7]. A crucial issue is *the choice of elements whose correspondence is sought*. In the wide-baseline set-up, local image deformation 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 main contribution of the paper is the proposal of two new types of distinguished regions together with efficient algorithms for their detection. Conceptually, these algorithms could be seen as processes that take the set of all subsets of the image pixels of all such shapes as input and select a subset possessing the distinguishing property. The art is in finding distinguishing properties that can be detected without the obviously prohibitive exhaustive enumeration of all subsets. For both new types of distinguished regions introduced, the *Separated Elementary Cycles of the Edge Graph (SECs)* and the *Maximally Stable Extremal Regions (MSERs)*, an efficient (near linear complexity) and practically fast (from fraction of a second to seconds) detection algorithm has been found. Low computational complexity and invariance to photometric and geometric transformation are desirable theoretical properties of the process of distinguished region detection. Stability, robustness and frequency of detection and hence usefulness 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 6 support the claim that the proposed DR types are very useful at least in man-made environments.

Reliable extraction of a manageable number of potentially corresponding image elements may be a necessary but certainly is 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 [6]. 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 [12, 19, 1, 20, 3, 17, 10, 9]. These methods exploit *local invariant descriptors* to limit the number of tentative correspondences. The key issues are 1. the choice of measurement regions, e.g. the parts of the image on which invariants are computed, and 2. the choice of invariants and 3. the method of selecting tentative correspondences given the invariant description. We discuss the structure of the class

of wide-baseline and recognition algorithms in Section 2 jointly with the review of the state-of-the-art. Such approach seems natural since differences in the published methods can be interpreted as a particular choice in one of the stages of a general framework.

Sections 3 and 4 give formal definitions of the two new types of distinguished regions, the Separated Elementary Cycles of the Edgel Graph and the Maximally Stable Extremal Regions. Both sections start with the definition of the underlying concepts. Next we present a detection algorithm, analyse its computational complexity and study invariance properties of the DR, concluding with remarks on robustness and relationship to other image processing methods. Examples of detected regions are shown later in the experimental Section 6. The SEC extraction algorithm operates on a novel representation of edge detector output called the Edgel Graph. Application of a commonly-used edge detector outputs typically a set of edgel strings. Without loss of efficiency, the modified linking stage produces a more structured and stable representation. As it is not in the main line of the paper, presentation of this minor contribution, perhaps of interest in its own right, was postponed till appendix A.

In Section 5 details of a novel matching algorithm (from the above-mentioned class) are given. A new *robust* approach is used for tentative correspondence computation. A robust similarity measure for comparison of local invariants replaces the common method based on Mahalanobis distance [14, 20, 15] which can be justified theoretically only under conditions that are almost certainly not met in the wide-baseline matching problem [5]. The robustness of 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.

Experimental results on outdoor and indoor images taken with an uncalibrated camera are presented in Section 6. On two simpler scenes, epipolar geometry is established using only a single type of distinguished regions. The potential for combination of multiple types of distinguished regions is demonstrated on perhaps the most difficult pair from the VALBONNE set. The last experiment can be viewed as a benchmark; results on the VALBONNE set has been presented in a number of papers on the topic [14, 12]. Presented experiments are summarised and the contributions of the paper are reviewed in Section 7.

2 Correspondence from Distinguished Regions

In the introduction, the concept of a distinguished region (DR) was described rather vaguely. In this section, we first present a formal definition of the DR concept, discuss some its properties and give examples of DR.

Definition 1 Distinguished region. *Let image I be a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Let $\mathcal{P} \subset 2^{\mathcal{D}}$, i.e. $\mathcal{P}$ is a subset of the power set (set of all subsets) of $\mathcal{D}$. Let $\mathcal{A} \subset \mathcal{P} \times \mathcal{P}$ be an adjacency relation on $\mathcal{P}$ and let $f : \mathcal{P} \rightarrow \mathcal{T}$ be any function defined on $\mathcal{P}$ with a totally ordered range $\mathcal{T}$. A region $\mathcal{Q} \in \mathcal{P}$ is distinguished with respect to function f iff $f(\mathcal{Q}) > f(\mathcal{Q}'), \forall(\mathcal{Q}, \mathcal{Q}') \in \mathcal{A}$.*

In order to be invariant to geometric transformations from a group G, the set $\mathcal{P}$ must be closed under action from G and the extremal property f must be preserved. As an example, let us view the Harris interest point detector, an operator commonly used in stereo matching, as a particular type of a distinguished region detector. The system of subsets $\mathcal{P}$ of the image domain $\mathcal{D}$ considered is the set of all circles with a fixed radius. The 'quality' function f assigns a positive real number to any element $\mathcal{Q}$ of $\mathcal{P}$, so $\mathcal{T} = R^{+0}$. The quality can be expressed as a function of the eigenvalues of the second moment matrix computed on $\mathcal{Q}$. The adjacency relation is defined the by maximum distance on centres of the circular regions that are subject to non-maxima suppression. Under translation and rotation, the set of all circles of a given radius is closed and since eigenvalues of the second moment matrix are rotationally and translationally invariant, the Harris operator detects the same distinguished regions under rigid transform. Under more complex geometric transformation (similarity, affinity) this is no more the case. The Maximally Stable Extremal Regions defined in Section 4 are an example of a distinguished region type invariant to a much broader class of geometric and photometric transforms. The invariant function f is typically constructed assuming local planarity and a continuous image domain and range. The practical value of a DR type given by stability w.r.t. viewpoint and illumination change must be established experimentally.

Note that we do not require DRs to have any transformation-invariant property that is unique or rare in the image. In other words, DRs need not be discriminative (salient). If a local frame of reference is defined on a DR by a transformation-invariant construction (projective, affine, similarity invariant), a DR may be characterised by invariant measurements computed on any part of an image specified in the local (DR-centric) frame of reference. We used the term **measurement region** for this part of the image.

Related work. Since the influential paper by Schmid and Mohr [15] many image matching and wide-baseline stereo algorithms have used Harris interest points as distinguished regions. Tell and Carlsson [17] proposed a method where line segments connecting Harris interest points form measurement regions. The MRs are characterised by scale invariant Fourier coefficients. 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 [10], 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 affine-invariant distinguished regions, one based on local intensity extrema [20] the other using point and curve features [19]. In the latter approach, DRs are characterised by measurements from the inside an ellipse, constructed in an affine invariant manner. Lowe [9] describes the 'Scale Invariant Feature Transform' approach which produces a scale and orientation-invariant characterisation of interest points.

So far we have focused on the selection of the elements to be put into correspondence (the DRs) and on the process of construction of measurement regions. Having two sets of DRs , how can the problem of epipolar geometry estimation by attacked? As mentioned in the introduction, in problems of realistic size it is clearly impossible to perform a brute-force search for the best globally consistent epipolar geometry. Instead, algorithms described in the literature have adopted strategies with a similar structure whose core is summarised in the following four steps:

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

1. Detect *distinguished regions*.
2. Describe DRs with invariants computed on measurement regions.
3. Establish tentative correspondences of DRs.
4. Estimate epipolar geometry in a hypothesise-verify loop.

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. The most simple situation arises if a local affine frame is defined on the DR. 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 [15, 14] or affine invariants must be used [20]. Selecting mutually nearest pairs in Mahalanobis distance is the most common method [14, 20, 15]. 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 [18].

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.

3 Elementary Separated Cycles of the Edgel Graph

In the literature on wide-baseline stereo, edge detectors have received significantly less attention than interest point operators. However, subsets of the edge map can provide, if extracted with good repeatability, richer geometric and photometric information than interest points. One such subset possessing properties desirable for discriminative regions is the set of *separated elementary cycles* (SECs) of the edgel graph. The novel concept is defined, together with other graph-theoretical concepts needed in the section, in Table 1. From this point it is assumed that the reader understands the concept of SEC and is familiar with edge detector output representation by the Edgel Graph. If in doubt, please re-read the definitions of Tab. 1 and look at the edgel graph and SEC visualisation shown in Figure 1. Details given in Appendix A describing the construction of the of the Edgel graph may be helpful too.

The motivation for investigating SECs is the following:

- We have observed that the number of elementary separated edge cycles is limited in most images. Even in textured ares the number of separated cycles is low (unlike the number of interest points), since typically non-separated cycles are formed.

- Geometric constraints stronger than single point correspondences can be obtained from the edgel strings associated with a separated cycle, e.g. local scale (area of the cycle) or even a full local affine reference frame (e.g centre of gravity with bitangents or other invariant points on the cycle). If more constraints are generated per DR, the number of correspondences defining uniquely the epipolar geometry is reduced. Consequently, estimation of epipolar geometry (e.g. by RANSAC) is either faster or feasible with a lower number of correct tentative correspondences.

- Invariants computed from the shape of the edgel string of a separated cycle can be exploited to reduce the number of tentative correspondences.

- The proposed approximate algorithm for SEC detection 2 guarantees that edgel strings of each cycle form a Jordan curve. Thus each cycle partitions the image plane into an 'inside' and 'outside'. The partitioning is preserved under practical perspective transforms. Measurements from the 'inside' can be therefore exploited in establishing tentative correspondences. The same is true for any measurement computed in a reference frame defined in an invariant manner on the edgel strings.

Problem 1: Invariance. Unlike the extremal regions described in 4, the employed Deriche edge detector [2] is not even scale invariant. In our experience (e.g. on the images presented in Section 6), a significant percentage of edgels is detectable over a range of scales. Only a few SEC are required to compute the epipolar geometry (depending

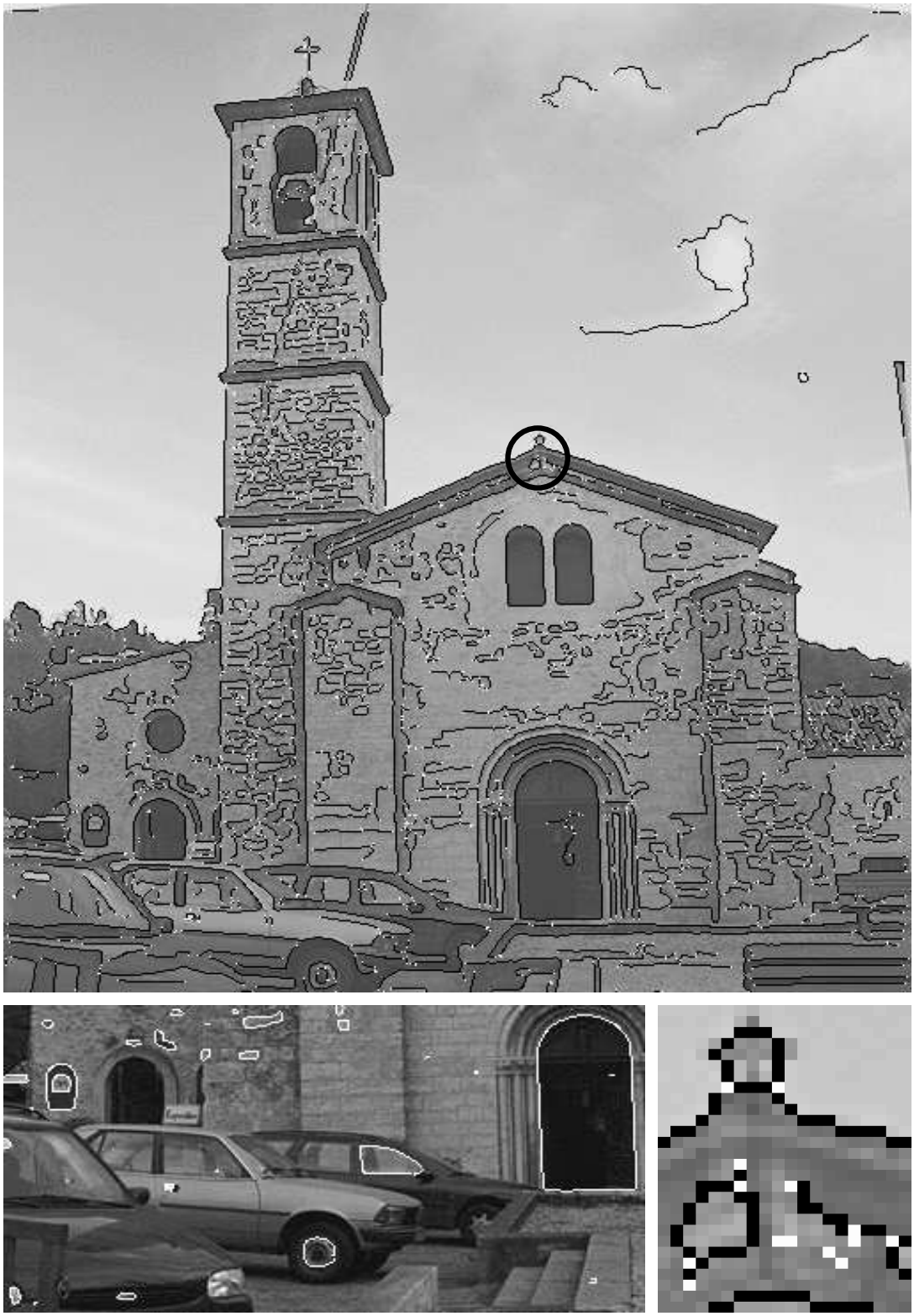


Figure 1: SEC detection on VALBONNE-003. The edgel graph $\mathcal{G}$ (top); each vertex of $\mathcal{G}$ is represented by a white point, each edge with the associated edgel strings (shown in black). The region in the black circle with two separated elementary cycles is magnified at bottom right. SECs detected in the lower left part of the image are depicted at bottom left.

Graph

$\mathcal{G}$ is a triplet (V, E, ε), where V is the set of vertices of $\mathcal{G}$, E the set of edges of $\mathcal{G}$, and $\varepsilon : E \to \binom{V}{2}$ is the graph adjacency function. A graph according to this definition is undirected and may have self-loops and multiple edges between vertices (such graphs are also called multigraphs).

Elementary Cycle

in graph $\mathcal{G}$ is a sequence $v_1, e_1, ..., v_n, e_n$, of vertices $v_i \in V$ and edges $e_i \in E$ without repetition such that each consecutive two vertices are adjacent and the last and the first vertices are adjacent, i.e $\varepsilon(e_i) = \{v_i, v_{i+1}\}, 1 \leq i < n$ and $\varepsilon(e_n) = \{v_n, v_1\}$.

Separated Elementary Cycle (SEC)

is an elementary cycle not sharing any edge with another elementary cycle.

Edgel String $\mathcal{S}$

is a connected set of edgels each having 2 neighbours. $\mathcal{S} \subset \mathbb{Z}^2$.

Start Edgel

is an edgel that has a number of neighbouring edgels different from 2, i.e. 0,1,3 or 4.

Edgel Graph

is an attributed (multi) graph $\mathcal{G} = (V, E, \varepsilon, A_v, A_e)$ representing the output of an edge detector. Each vertex $v \in V$ represents a Start Edgel. The vertex attribute function $A_v{:}\, V \to \mathbb{Z}^2$ associates position of the edgel with the node. Each $e \in E$ represents an Edgel String. The edge attribute function $A_e{:}\, e \to L \subset \mathbb{Z}^2$ associates an Edgel String with each edge.

Table 1: **Definitions** used in Section 3

on what geometric constraints are associated with each correspondence of cycles). In the context of our application, it is not necessary to 'interpret' the edges. Edges arising from scratches on a surface, albedo change or surface orientation discontinuity are equally useful, as long as they are repeatedly detected. Even shadows are helpful if they are present in both images. This is in contrast to some traditional use of the edge detector (interpretation of edges as surface discontinuities, association of edges with primitives of line drawings) where such edges would be considered spurious.

Problem 2: Computational complexity. The problem of *enumeration of all elementary cycles of a graph* has been studied in combinatorial mathematics. The bound on its time complexity is given in [13] as $O(Nm + n + m)$, where n is the number of vertices, m is the number of edges N is the size of the output, i.e. the number of cycles. In our application, the Nm term is prohibitive. In a complex image like VALBONNE-003 (Fig. 1) the number m of edgel strings (not edgels!) is above 3000.

Enumeration of separated elementary cycles. For a small graph, the problem becomes computationally tractable if only *separated* elementary cycles are required. Since each edge can only belong to at most one SEC, the size of the output $N < m$ and the bound on the complexity becomes $O(m^2)$. This is still not practical for our application. We therefore propose an approximate algorithm that is fast and simple[1] and has linear time complexity $O(\max(n, m))$, where n is the number of nodes and m the number of edges. The structure of the algorithm is shown below.

Algorithm 2: Approximate enumeration of separated elementary cycles

Input: undirected graph $\mathcal{G} = (V, E, \varepsilon)$

Output: list of elementary separated cycles

1. do 2 times (or until not new self-loops found; see text)
2. Remove vertices of degree 1.
3. Remove degree 2 nodes that are not self-loops, propagating edgel strings.
4. Add to output all edges that are self-loops and remove them from $\mathcal{G}$.

Before describing in detail steps 1-4, let us first explain the nature of the approximation of Algorithm 2. We detect only those loops that can be reduced by steps 2 and 3 to a single self-loop edge. In the first step, all vertices of degree 1 are removed since they cannot be part of any cycle. This process can be completed in a single sweep through the list of vertices. The operation has complexity linear in the number of vertices, since each vertex is considered at most twice. The second 'touch' of the vertex happens if

[1] On a the VALBONNE-003 image (see Fig. 1) the time to compute all SECs is approximately 0.2 seconds on a SPARC ultra. Implemented in approximately 60 lines of C code (using a good graph library).

it is (the single vertex) adjacent to a vertex of degree 1. What is left are separated cycles, complex cycles and connections between them. Most separated cycles become sequences of vertices of degree 2. These sequences are reduced in step 2 to single vertex with a self-loop. All degree 2 nodes without a self-loop are deleted from the graph; before removing an edge, the edgel string associated with it is moved to an adjacent edge. Steps 2 and 3 are implemented as a single sweep trough the list of vertices and edges respectively. Steps 2-4 must be in general iterated, since detection of a cycle and a subsequent removal of a self-loop may transform a vertex of degree 3 to degree 1. In experiments, this never happened after the second iteration. However, we have constructed examples where the number of iterations needed is $\log m$. Facing diminishing returns, we set $k = 2$ to maintain the linearity of the algorithm.

Robustness. The concept of an elementary cycle is very brittle. A single one-pixel gap — a common event especially in areas of complex intensity structure — will break a cycle. To counter the problem, edges are inserted into the graph between two degree one vertices, if they correspond to edgels satisfying a distance constraint. The distance threshold and the edge detector filter width are the only parameters of the method.

4 Maximally Stable Extremal Regions

In this section, we propose a class of distinguished regions that is based solely on an extremal property of the intensity function in the region and on its outer boundary. The so called *Maximally Stable Extremal Regions (MSERs)* can be defined on any image (even high-dimensional) whose pixel values are from a totally ordered set. The formal definition of the MSER concept and the necessary auxiliary definitions are given in Table 2.

The concept can be explained informally as follows. Imagine all possible thresholdings of an input gray-level image I, say with a common range $\mathcal{S} = \{0, 1, \ldots, 255\}$. 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 union of all connected components of all frames of the movie is identical to the set of all maximal regions; minimal regions could be obtained by inverting the intensity of I and running the same process. On many images one observes that local binarisation is stable over a large range of thresholds in certain regions. Such regions are of interest since they posses the following properties:

- **Invariance to monotonic transformation** $M : \mathcal{S} \to \mathcal{S}$ of image intensities.

  The set of extremal regions is unchanged after transformation M, $I(p) < I(q) \to M(I(p)) = I'(p) < I'(q) = M(I(q))$ since M does not affect adjacency (and thus contiguity) and intensity ordering is preserved.

Image

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

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 $\sum_{i=1}^{n} |p_i - q_i| \leq 1$.

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 region $\mathcal{Q}$ 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). Δ is a parameter of the method.

Table 2: **Definitions** used in Section 4

- **Invariance 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** $O(n \log \log n)$, i.e. almost in linear time for 8 bit images.

Due to lack of space, the algorithm for extremal region detection can be only briefly outlined.

Algorithm 3: Enumeration of Extremal Regions. (outline)

Input: Image I

Output: list of nested extremal regions

1. For all pixels sorted by intensity
2. Place pixel in the image.
3. Update the connected component structure.
4. Update the area for the effected connected component.
5. For all connected components
6. Local minima of the rate of change of its area define stable thresholds.

The computational complexity of step.1 is $\mathcal{O}(n)$ if the image range $\mathcal{S}$ is small, e.g. the typical $\{0, \ldots, 255\}$, and sorting can be implemented as BINSORT [16]. As pixels ordered by intensity are placed in the image (either in decreasing or increasing order), the list of connected components and their areas is maintained using the efficient union-find algorithm [16]. The complexity of the algorithm is $\mathcal{O}(n \log \log n)$. The process produces a data structure holding the area of each connected component as a function of a threshold. A merge of two components is viewed as the end of existence of the smaller component and the 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. In the output, each MSER is represented by a local intensity minimum (or maximum) and a threshold.

Notes. The structure of algorithm 3 and an efficient **watershed algorithm** [21] is essentially identical. However, the structure of *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 thresholds were regions merge (and watershed basins touch). Such threshold are of little interest to us, 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 thresholded image. However, no global or 'optimal' threshold is sought, all thresholds are tested and the stability of the connected components evaluated. Finally, the output 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.

5 Matching

The most common method for establishing tentative correspondence is based on Mahalanobis distance (MD) [14, 20, 15]. However, the method can be criticised on both theoretical and practical grounds. In the stereo matching problem, no training samples are available and the use of Mahalanobis distance is equivalent to whitening the *total* covariance matrix and computing the Euclidean distance. No attempt is made to estimate the within-class covariance matrix (the covariance of the errors in corresponding measurements in the two images) nor the between-class covariance matrix. This is equivalent to the assumption that two covariances are equal [5] which in most problems is far from true. Concerns about the inherent Gaussian assumption may be voiced too. From the practical point of view, MD is not robust – a single 'wild' measurement can make it arbitrarily large. Our experiments have shown that often at least *some* of the affine invariants used are unstable.

On the other hand, the robustness of 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 or completely wrong. The former helps establishing reliable tentative correspondences, the influence of the latter is limited by robustness of the approach. We first define the similarity measure and then briefly comment on its properties.

Each DR is described by a measurement vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$. In the matching problem there are two sets $\mathcal{L}$ and $\mathcal{R}$ of DR measurement vectors originating from the 'left' and 'right" image respectively. The task is to find tentative matches given the local description. The set of initial correspondences is formed as follows. Two regions with descriptions $\mathbf{x} \in \mathcal{L}$ and $\mathbf{y} \in \mathcal{R}$ are taken as a candidates for a match iff $\mathbf{x}$ is the most similar measurement to $\mathbf{y}$ and *vice-versa*, i.e.

$$\forall \mathbf{x}' \in \mathcal{L} \setminus \mathbf{x} : d(\mathbf{x}, \mathbf{y}) < d(\mathbf{x}', \mathbf{y}) \qquad \text{and} \qquad \forall \mathbf{y}' \in \mathcal{R} \setminus \mathbf{y} : d(\mathbf{y}, \mathbf{x}) < d(\mathbf{y}', \mathbf{x}),$$

where d is the asymmetric similarity measures defined below. In the computation of $d(\mathbf{x}, \mathbf{y})$ each component of the measurement vector is treated independently. The similarity between the i-th component of $\mathbf{x}$ and $\mathbf{y}$ is measured by the number of vectors $\mathbf{y}'$ whose i-th measurement is closer. In other words the similarity in the i-th component is the rank of the measurement from $\mathbf{y}$ among all measurements $\mathbf{y}'$ from $\mathcal{R}$:

$$\text{rank}^i_{\mathbf{x},\mathbf{y}} = \text{card}(\{\mathbf{x}' \in \mathcal{L} : |x'_i - y_i| \leq |x_i - y_i|\}). \quad (1)$$

The overall similarity measure is then defined as follows

$$d(\mathbf{x}, \mathbf{y}) = \text{card}(\{i \in \{1, \ldots, n\} : \text{rank}^i_{\mathbf{x},\mathbf{y}} < t\}), \quad (2)$$

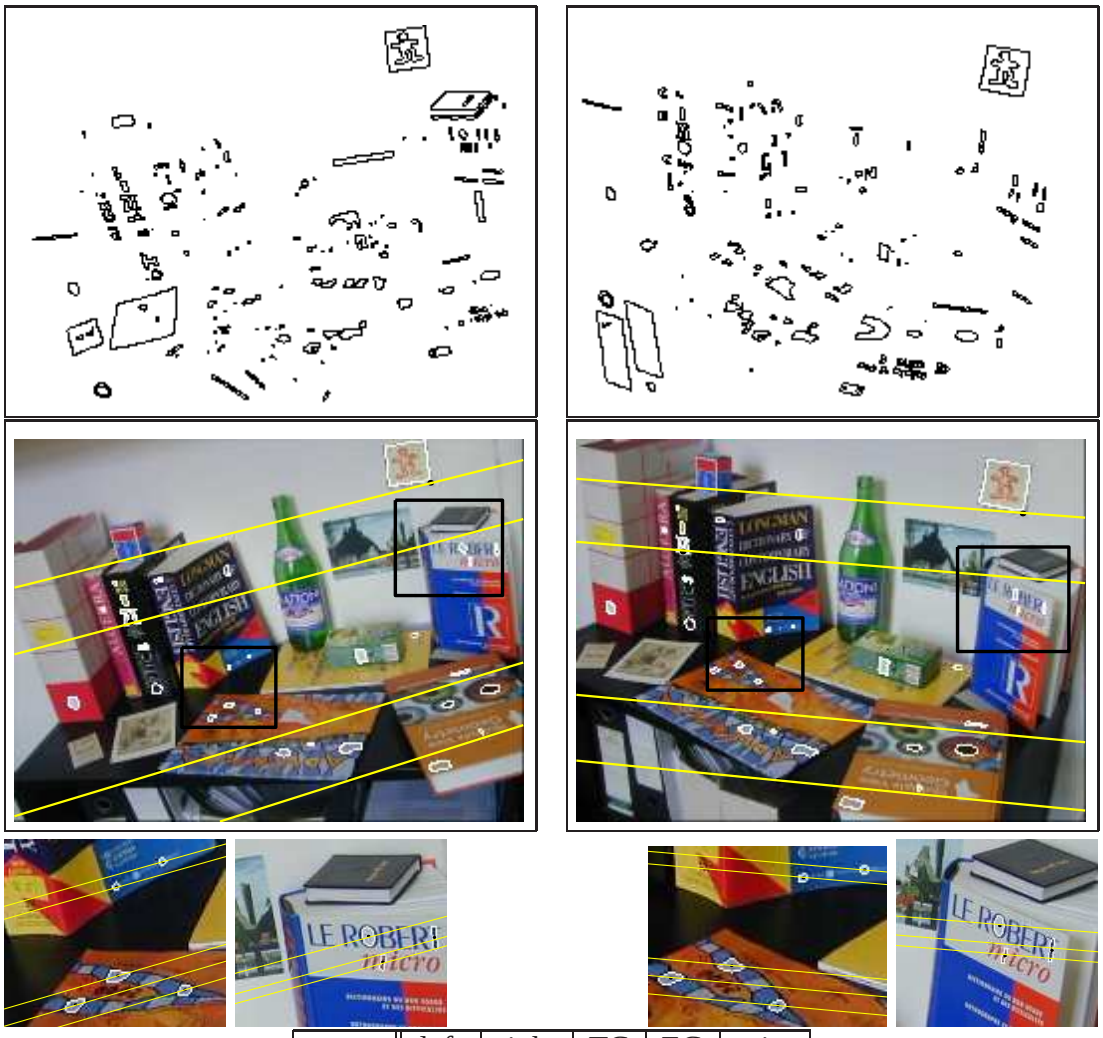
where n is dimension of the measurements vector and t a predefined ranking threshold. The computation of $d(\mathbf{y}, \mathbf{x})$ is analogous with the roles of $\mathcal{L}$ and $\mathcal{R}$ interchanged. The most important property of d is that the influence of any single measurement is limited to 1. Only the main idea of the probabilistic error model behind the design may be mentioned due to limited space. Under a very broad range of error models, corresponding measurements are more likely to be below the ranking threshold than a mismatch. One of the interesting properties of the similarity measure is its invariance to any monotonic transformation applied to elements of the measurement vector $\mathbf{x}$, so measurement of different orders of magnitude are easily handled.

6 Experiments

In all experiments, the following parameters of the matching algorithm were used. Each measurement region was described by 21 general colour moment invariants of Midru et. al. [11] computed from four measurement regions (MRs). The MRs defined in terms of affine-invariant constructions on the DR boundaries were the following: the DR itself and its convex hull scaled by factors of 1.5, 2 and 3. The rank threshold, a parameter of the robust similarity measure, was set to 7. Tentative correspondences comprised only those pairs whose colour invariants were mutually nearest in the robust similarity measure. Note that the similarity was computed in 84-dimensional space (21 invariants, 4 MRs). Epipolar geometry was estimated by the 7-point algorithm [7]. In all experiments, only a linear algorithm is used [7] to estimate epipolar geometry; no effort was made to improve the precision by known methods such as bundle adjustment, correlation, or homography growing.

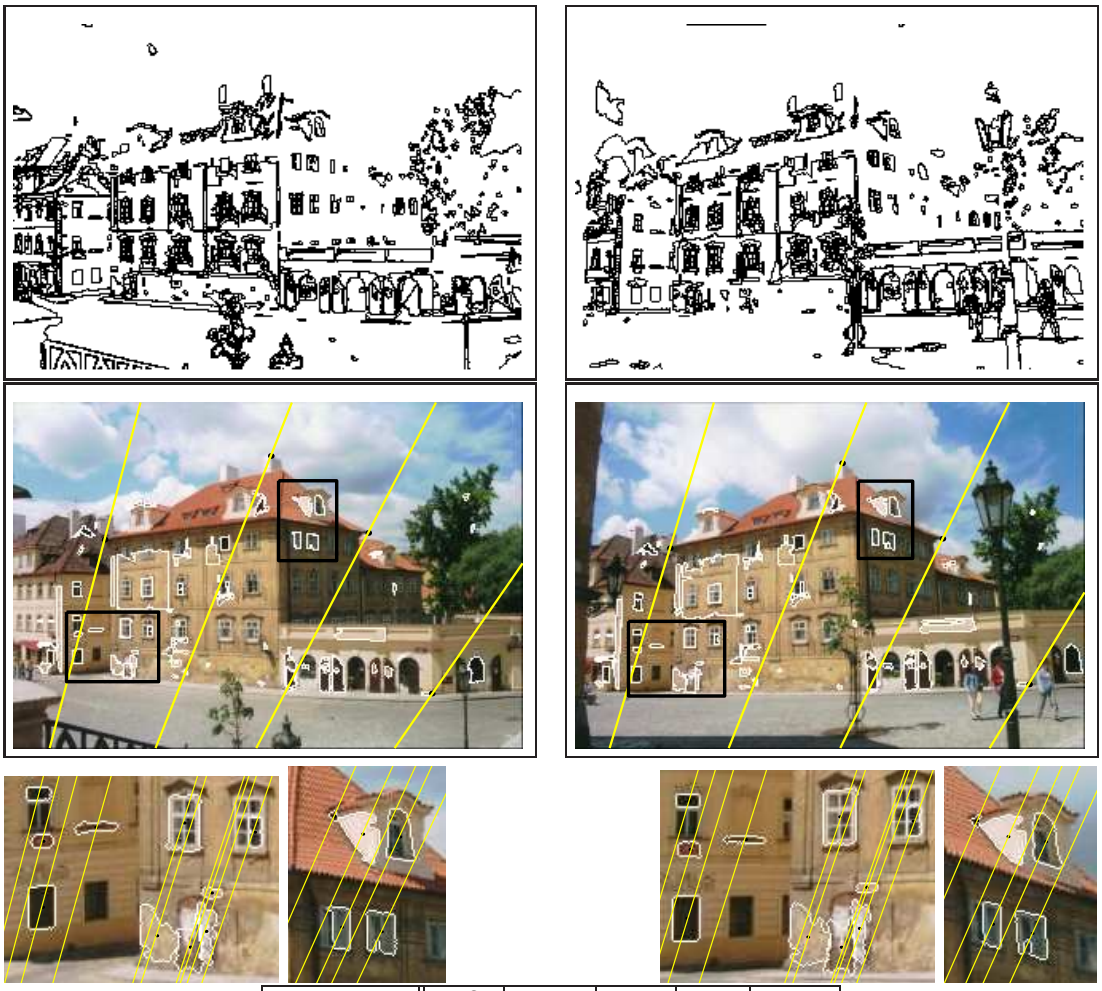
Experiment I.: Epipolar geometry from SEC correspondences. The potential of the Separated Elementary Cycles for wide-baseline stereo was evaluated on images of an office scene. The BOOKSHELF dataset[9] contains approximately ten images taken from significantly different viewpoints. The density, precision and repeatability of the SEC output is shown on the pair images depicted in Figure 2 (middle row); more successful wide-baseline matching experiments are reported in [8].

[9]The data will be made publicly available.



	left	right	TC	EG	miss
SECs	178	162	52	29	1

Figure 2: BOOKSHELF: SECs detected in a pair of images (top row), estimated epipolar geometry (middle row) and a close-up of selected areas marked with black rectangles in the originals (bottom row).



	left	right	TC	EG	miss
MSERs -	558	437	143	41	0
MSERs +	404	467	120	37	0
total	962	904	263	78	0

Figure 3: KAMPA (easy): MSERs (both intensity minima and maxima) detected in a pair of images (top row), estimated epipolar geometry (middle row) and a close-up of selected areas marked with black rectangles in the originals (bottom row).

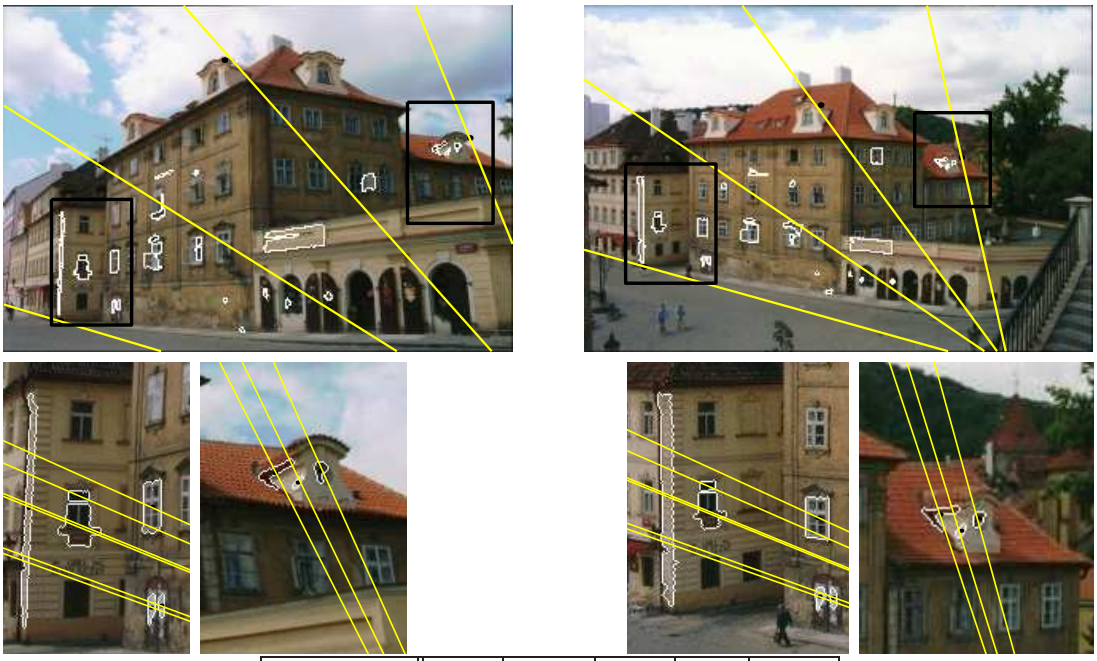
The top row of Figure 2 shows edgel strings that are attributes of edges of separated elementary cycles of the Edgel Graph. The number of SECs in the left and right images was 178 and 162 respectively. In the middle row, the input images with overlayed epipolar lines are shown. The SECs that formed tentative correspondences and are consistent with the estimated epipolar geometry are highlighted. The number of SECs with mutually nearest invariant descriptions, i.e. the number of tentative correspondences, was 52 in this test. The RANSAC procedure found an epipolar geometry consistent with 29 tentative correspondences of which 28 are correct. The numbers of detected SECs in the left and right images, tentative correspondences (TC), epipolar geometry consistent correspondences (EG) and the number of mismatches (miss) are summarised in the caption of Figure 2. Mismatches are correspondences consistent with the estimated epipolar geometry that are not projections of the same part of the scene. The ratio TC/EG determines the average number of RANSAC hypothesis-verify attempts[2] and hence the speed of epipolar geometry estimation.

The bottom row of Figure 2 shows close-ups of two rectangular regions selected from the left and right images respectively. The subimage in the bottom left corner is interesting because the three SECs have identical light blue insides and almost affinely equivalent shape. The SEC regions formed correct tentative correspondence probably because of the differences in the larger and discriminative measurement regions.

Experiment II.: Epipolar geometry from MSER correspondences. Maximally Stable Extremal Regions were evaluated on images of an urban scene. Images from the KAMPA dataset[9] are shown in Figures 3 and 4. The stereo problem is much more difficult for the pair presented in Figure 4, where the viewpoint change induces significant perspective effects and change of scale. Moreover, contrast is very low on the right side of the images. Taking into account the changing skies, the part of the scene visible in both images covers less than 50% of the images. The extent of local changes is clearly visible in the close-ups presented in the bottom row of Figure 4. As an example, consider the change near the attic window shown in the bottom right image. The window is viewed from a very different angles; the background changes dramatically. The repeatability of the MSER around the window is surprisingly good, despite the acute viewing angle and small resolution. The close-up at bottom left demonstrates that despite the large change in viewpoint, certain regions remain remarkably stable.

Compared with the previous example, the simpler stereo problem presented in Figure 3 does not say much about the limits of MSER-based matching. The image pair is included mainly to demonstrate the high density of MSERs from which tentative correspondences were formed. The number of MSERs detected in the left and right images (top, Figure 3) is above 900. In fact, two types of MSERs were extracted; those corresponding to local intensity maxima (MSER+) and to local minima respectively (MSER-). In total, 263 tentative correspondences with mutually nearest invariant descriptions are input into RANSAC. Since contrast reversal is not expected, only correspondence within the respective classes (either '+' or '-') are allowed. The RANSAC

[2] the average number of trials is approximately $(\mathrm{TC}/mboxEG)^7$)



	left	right	TC	EG	miss
MSERs -	292	413	72	12	3
MSERs +	277	414	73	11	1
total	569	827	145	23	4

Figure 4: KAMPA (hard): Estimated epipolar geometry (top row) and a close-up of selected areas marked with black rectangles in the originals (bottom row).

procedure found an epipolar geometry consistent with 78 tentative correspondences and no mismatches. The number of MSERs detected in the left and right images, the number of tentative correspondences (TC), epipolar geometry consistent correspondences (EG) and mismatches (miss) are summarised in the caption of Figure 3. In the top row if Figure 3 we tried to visualise the MSERs. It is not easy, since MSER do not form a partitioning but rather a tree of nested regions; it is not possible to trace boundaries of individual MSERs in a binary image. The MSER boundary image at least clearly shows that the density of MSERs is high almost everywhere, the sky and the featureless road being the exceptions.

Returning to the more difficult problem of Figure 4, we see that the number of tentative correspondences is much lower – 145. Using RANSAC, an epipolar geometry was found that was consistent with only 23 tentative correspondences of which 4 are mismatches. Both the small number of epipolar geometry consistent correspondence and the small ratio of EG consistent to tentative correspondences suggests we are near the limits of the method.

Experiment III.: Cooperation of multiple DR detectors. This experiment was conducted on one of the most difficult pairs from the VALBONNE set, see Figure 5.

This matching problem can be viewed as a benchmark since results on VALBONNE images have been published in the literature [14, 12]. Several factors contribute to the complexity of this matching problem. The main problem difficulty is that only a set of parallel planes with small relative depth is visible in both images. Importantly, not a single wall with significant relative depth is present in both images (the walls perpendicular to plane of the portal). For example, only the frontal side of the bell tower is visible in both images.

We have tacitly used combination of two types of DR detectors, the MSER+ and MSER- , already in Experiment II. Here we combine SECs with the two types of MSERs. The number of DRs detected in each of the images is fairly high. The fraction of DRs that gave rise to a tentative correspondence is very low. This is due to the non-distinctive nature of DRs on the wall; often a DR covered a single stone. Perhaps surprisingly, some of the DR corresponding to a single stone were successfully matched, probably because of the fairly large measurement regions combined in the robust similarity measure. The numbers of DRs detected in the left and right image, the number of tentative correspondences (TC), epipolar geometry consistent correspondences (EG) and mismatches (miss) are summarised in the caption of Figure 5. The table shows that the number of correct matches consistent with the found epipolar geometry for each type of DRs was between 8 and 10, which is insufficient for reliable EG estimation since. In total, 26 correct correspondences are found, which is much less likely to arise randomly.

7 Conclusions

In the paper we first discussed the choice of image elements that are put into correspondence in the wide-baseline matching problem. We introduced and defined formally the concept of a distinguished region and we argued they are eligible candidates for matching.

The main contribution of the paper is the introduction of two new types of distinguished regions. For both types, the *Separated Elementary Cycles of the Edge Graph (SECs)* and the *Maximally Stable Extremal Regions (MSERs)*, an efficient (near linear complexity) and practically fast detection algorithm was presented. Experimentally we showed the stability of the proposed DRs in disparate views of real-world scenes with significant change of scale, camera rotation, and 3D translation of the viewpoint.

In a second contribution, a robust similarity measure for establishing tentative correspondences was proposed. 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 distinguished region.

Good estimates of epipolar geometry were obtained on challenging wide-baseline problems with the robustified matching algorithm operating on the output produced by the proposed detectors of distinguished regions. Fully affine distortions and significant occlusion were present in the tests. Test images included both outdoor and indoor

scenes, some already used in published work.

The use of multiple types of distinguished regions was demonstrated in an experiment conducted on a non-trivial pair from the VALBONNE set, where the estimation process of epipolar geometry via RANSAC failed for any single DR type. Finally, we proposed a modification of the linking part of an edge detector increasing its repeatability.

A Constructing the edgel graph $\mathcal{G}$

First, the Deriche filter is applied [2] and non-maxima suppression is carried out. Unlike in standard edge detectors, the output of hysteresis thresholding is not a set of edgel strings but a graph. An *edgel* is a pixel, i.e. an element of $\mathbb{Z}^2$, that was assigned value 1 in the hysteresis thresholding process. Definitions for *edgel string* and *start (end) edgel* are given in Table 1. The edge detector output[3] is transformed in an attributed graph $\mathcal{G} = (V, E, \varepsilon)$. The the set of vertices V and the set of edges E are extracted in the following manner. Each start (end) edgel is represented by a vertex, each edgel string is represented by an edge between vertices of the start and end edgels connected by the edgel string. The edgel string (represented as a list of edgels) is stored as an attribute of the edge it gave rise to.

Technical details. The algorithm for construction of $\mathcal{G}$ as described above of would not be able to represent edgel strings that do not have start and end edgels (closed loops of edgels; not to be confused with cycles of $\mathcal{G}$!). The situation is easily detected, an arbitrary edgel is chosen as the start (and end) edgel and a vertex representing it is inserted into the graph. The graph may have vertices of degree zero (representing isolated edgels) as well as self-loops, i.e. edges starting and ending in the same vertex, representing closed edgel strings. So far the type of neighbourhood relationship has not been specified. To interpret the edgels as strings as far as possible a mixture of 4 and 8 connectivity is used. We start by checking the number of 4-neighbours of an edgel. If two or more neighbours are found, the search for neighbours stop. If less than two are found, a subset of 8-neighbours consistent with the 4-neighbours is checked. The switching between 4 and 8 connectivity is clearly apparent in the scaled-up part of the VALBONNE-003 image shown at the bottom right of Figure 1. Edgel strings are in black, each white point will give rise to a vertex in $\mathcal{G}$. Each connected component of black points is represented as an edge in $\mathcal{G}$. The complexity (linear in the number of pixels) as well as practical speed of construction of $\mathcal{G}$ is effectively identical to a standard edge linking procedure. In common implementations, edgel strings are either broken or follow a random path at points where there are more than two neighbours. Together with the hysteresis thresholding process, both common methods significantly reduce repeatability of the edge detector output, since

[3] It would be more appropriate to speak about edgel detector in the context of this section. The term 'edge' refers here to an entity in the graph $\mathcal{G}$

pixels above the higher threshold are connected with different subset of edgels each time.

References

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

[2] R. Deriche. Using Canny's criteria to derive a recursively implemented optimal edge detector. *IJCV*, 1:167, 1987.

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

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

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

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

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

[8] hidden to conceal author's identity. .

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

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

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

[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] R.C Read and R.E Tarjan. Bounds on backtrack algorithms for listing cycles, paths, and spanning trees. *Networks*, 5:237–252, 1975.

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

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

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

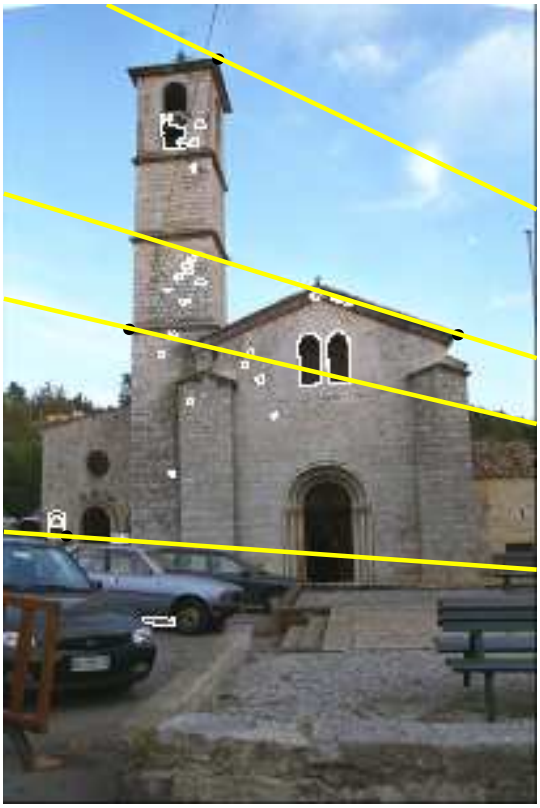
[17] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV00*, 2000.

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

[19] 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.

[20] 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.

[21] 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.



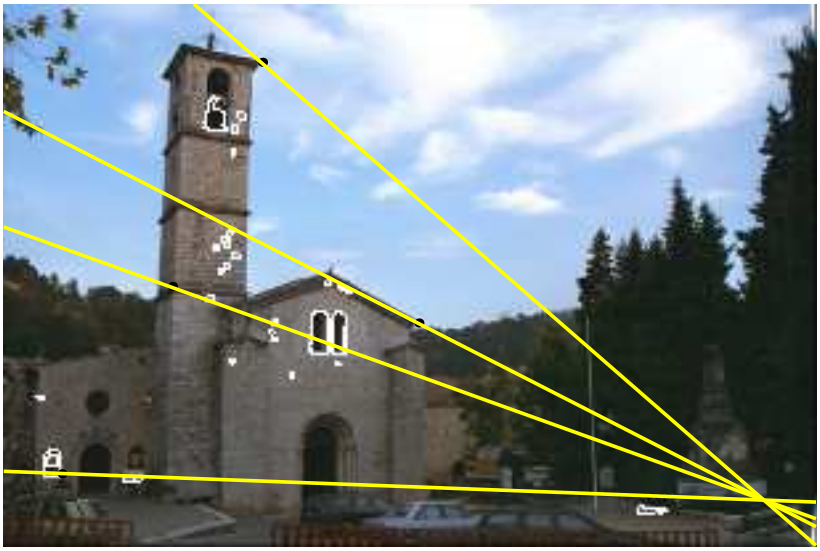


Figure 5: VALBONNE: Estimated epipolar geometry.

	left	right	TC	EG	miss
SECs	529	523	50	11	3
MSERs -	320	127	49	8	0
MSERs +	518	362	89	11	1
total	1367	1012	188	30	4