Shape Guided Maximally Stable Extremal Region (MSER) Tracking

Michael Donoser, Hayko Riemenschneider and Horst Bischof*
Institute for Computer Graphics and Vision, Graz University of Technology
{*donoser,hayko,bischof*}*@icg.tugraz.at*

Abstract

Maximally Stable Extremal Regions (MSERs) are one of the most prominent interest region detectors in computer vision due to their powerful properties and low computational demands. In general MSERs are detected in single images, but given image sequences as input, the repeatability of MSER detection can be improved by exploiting correspondences between subsequent frames by feature based analysis. Such an approach fails during fast movements, in heavily cluttered scenes and in images containing several similar sized regions because of the simple feature based analysis. In this paper we propose an extension of MSER tracking by considering shape similarity as strong cue for defining the frame-to-frame correspondences. Efficient calculation of shape similarity scores ensures that real-time capability is maintained. Experimental evaluation demonstrates improved repeatability and an application for tracking weakly textured, planar objects.

1 Introduction

The detection of interest points and local features constitutes the basis for many important computer vision tasks. For example, object recognition, stereo matching, image mosaicking, robot navigation, etc. rely on the detection of interest points which possess some distinguishing, highly invariant and stable properties. Such structures provide a compact and abstract representation of patterns in an image.

Numerous interest point detection algorithms have been proposed in the recent years returning structures like corners [8], blobs [13] or edges [14]. Detailed evaluations and comparisons of different interest point detectors are available, e. g. by Mikolajczyk et al. [9]. These evaluations revealed that the Maximally Stable Extremal Region (MSER) detector proposed by Matas et al. [7] performs best on a wide range of different test sets. An MSER is a distinguished region defined by an extremal property of its intensity function in the region and on its outer boundary. MSERs have all the properties required of a robust local detector.

If a sequence of images is available for interest point detection, temporal information can be included to improve the overall detection repeatability. For example, Video Google [12] describes an approach to object and scene retrieval based on tracked distinguished regions, where tracking and interest point detection are realized by different algorithms. Obviously, results would be improved if both detection and tracking would be based on the same principles.

Improving repeatability of MSER detection in video sequences was first addressed by Donoser and Bischof [2]. They exploited available data from the detection process in frame t to find corresponding regions in the next frame and demonstrated improved repeatability on several sequences. For finding corresponding regions, simple features like size, elongation and intensity values were used. Such an approach fails during fast movements, in cluttered scenes and when several similar sized regions appear close to each other as frequently occurring in man-made objects like signs, logos or inscriptions. For example Nister and Stewenius [10] proposed an efficient vocabulary tree structure for e. g. recognizing CD covers in a real-time demo. Their method utilizes MSERs as underlying interest region detector and would certainly benefit from more repeatable detection results. CD covers often contain similar regions of same size and appearance and therefore the standard MSER tracking method [2] often fails in this scenario.

In this paper we propose an extension to MSER tracking, where we use an efficient shape matching method to identify correspondences in a more robust way. Using shape similarity maintains the same repeatability on standard sequences, whereas much more

*This work was supported by the Austrian Research Promotion Agency (FFG) project FIT-IT CityFit (815971/14472-GLE/ROD) and the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.444

robust results are achieved in more complex scenarios. Since a large number of shape comparisons have to be performed for finding the correspondences, high efficiency is required for this step. Fortunately, recent progress in the area of shape matching enables comparisons within milliseconds.

The outline of the paper is as follows. Section 2 summarizes the proposed shape guided MSER tracking method. It outlines the basics of MSER detection and tracking and shows how to integrate an efficient partial shape matcher to define frame-to-frame correspondences. Section 3 compares our proposed method to the original version demonstrating improved repeatability and an application for tracking weakly textured, planar objects.

2 MSER tracking using shape

This section introduces our shape guided MSER tracking method. We first summarize the properties of the MSER detector in Section 2.1. Section 2.2 describes the extension to MSER tracking, outlining the required region feature comparison, which is used to find the frame-to-frame correspondences. Finally, in Section 2.3 we show how an efficient shape matching method is integrated for improving detection repeatability.

2.1 MSER detection

Maximally Stable Extremal Region (MSER) detection was proposed by Matas et al. [7]. It detects a set of connected regions from an image, where each region is defined by an extremal property of the intensity function within the region to the values on its outer boundary. MSERs are invariant to continuous geometric transformations and affine intensity changes and are detected at several scales. MSER are further considered as the fastest interest point detection method, since algorithms for calculating MSERs in linear time [11] are available.

MSERs are detected by analyzing a unique grayscale image representation denoted as component tree. The component tree C can be build for every image with pixel values coming from a totally ordered set, e. g. from a standard grayscale image. Each node of the tree C contains a single connected region R_i, that is found as connected component (a so called extremal region) within binary threshold results $T_d = I_{in} \geq d$ of the input image I_{in}. By thresholding the image at all possible values from 255 down to 0, the component tree structure is built. Since regions in the tree can only become larger with decreasing threshold, this implicit inclusion relationship defines the edges of the component tree C. Therefore, a region R_i that is the father of a region R_j contains all the pixels of R_j and the root node contains all image pixels.

MSERs are selected nodes within the component tree, namely the most stable ones. For every node a stability value is calculated, which estimates the stability of region size over Δ levels of the component tree, where Δ is a fixed parameter of the method. The locally most stable ones are returned as the MSER detection result.

2.2 MSER tracking

The component tree is an effective data structure for detection of MSERs in single images and in addition constitutes the basis for the extension to robust tracking of MSERs. In general, if MSER detection results are required for a sequence of images, detection is done independently on every image. MSER tracking as proposed in [2] additionally integrates temporal information into the framework which allows to significantly reduce the computation time and to improve the repeatability of the detection results. The tracking algorithm starts with the analysis of the entire image I_t at frame t which results in a detection of MSERs for this image. Then each detected MSER of image I_t is tracked independently of all the others by performing three steps. First, a region of interest (ROI) of pre-defined size, centered on the center of mass of the MSER to be tracked, is propagated to the next frame. Second, the component tree is built for this ROI. Finally, the component tree is analyzed and the node which best fits to the input MSER is chosen as the tracked representation. In contrast to single image based MSER detection, in tracking every extremal region of the component tree is considered as potential correspondence (not only the maximally stable ones), which is the reason for the improved repeatability.

In order to identify the best fit M^*_{t+1} to the input MSER M_t, F-dimensional feature vectors $\mathbf{f} = (f_1, f_2, \dots f_F)$ are calculated for each of the N extremal regions R^j_{t+1} of the component tree C. The features calculated are region size, mean and minimum gray value, width and height of the bounding box, center of mass and region stability. The tracked representation M^*_{t+1} is then chosen as the one with the lowest weighted Euclidean distance between its feature vector and the vector of the region to be tracked M_t by

$$M^*_{t+1} = \underset{R^j_{t+1} \in C}{\arg\min} \left[\sum_{i=1}^{F} \omega_i \left(f_i(R^j_{t+1}) - f_i(M_t) \right)^2 \right], \tag{1}$$

where ω_i defines a weight for each feature which can be used to adapt to different kinds of input data.

Obviously, feature comparison fails when the considered regions are too similar concerning the simple features. Therefore, we propose to replace this feature based comparison by analysis of the shape of the outer regpon contours as it is described in the next section.

2.3 Shape guided MSER tracking

The underlying idea of our approach is to use the shape of the MSERs to find the correspondences, since for every node of the tree the closed outer contour is provided for free during computation of the component tree. In general, the shape of the contour seems to be a much more discriminative feature than the ones used in [2]. Therefore, the entire MSER tracking method as described in the last section remains the same, we only replace the feature comparison with a score representing the shape similarity between regions.

Shape is the key feature in versatile applications of computer vision and therefore shape matching is a well investigated problem, where many different powerful solutions exist [1, 6, 4]. Since for our tracking application a large number of matches have to be performed per frame, most of these state-of-the-art approaches cannot be applied since they require several hundreds of milliseconds per single match. For example, the top performing shape matching method on the well known MPEG - 7 data set [4] requires half a second per match, which is not feasible in our scenario.

We propose to use a recent shape matching method denoted as IS-Match [3] for this purpose. This method uses chord angle based descriptors for solving an order preserving assignment problem. Integral images are used as efficient data structure enabling partial matching within a few milliseconds. Since sampled contour points are used as shape representation, point correspondences and a shape similarity score are provided per match.

We adapt the method of [3] to our purposes by precalculating the shape descriptor of our MSER to be tracked M_t, and by using a standard Procrustes distance to define the similarity score. We define the correspondence by returning the extremal region with the most similar shape, i. e. the region with the lowest Procrustes distance P_D. Therefore, we replace the feature comparison defined in Equation 1 with

$$M_{t+1}^{*} = \underset{R_{t+1}^{j} \in C}{\arg\min} \left[P_D \left(M_t, R_{t+1}^{j} \right) \right], \qquad (2)$$

where $P_D\left(R_1, R_2\right)$ is the estimated IS-Match Procrustes distance between the two closed contours R_1 and R_2. Please note, that matching is only performed for regions with approximately similar basic features.

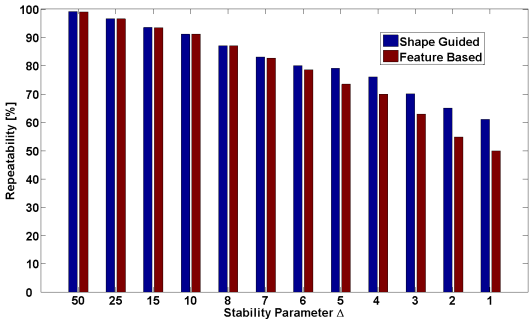

Figure 2. Comparison of repeatability scores for feature based and shape guided MSER tracking for different parameter values Δ.

Thus, in contrast to the original MSER tracking formulation, we do not use the simple features for finding the best correspondence, we solely use them for getting some candidates from the component tree, which are then validated by the accurate shape matching.

3 Experiments

As a first experiment we demonstrate our proposed shape guided MSER tracking method for the task of tracking weakly textured, planar objects through video sequences. We initialize the tracker in the first frame by drawing a bounding box over the object, and use the detected MSERs within the bounding box as initialization. We apply our method on sequences of a recently proposed framework for comparing tracking results [5] and selected frames are shown in Figure 1. As can be seen each of the initialized regions is robustly tracked through the entire sequence yielding a 100% repeatability score. Please note, that the obtained correspondences provided by shape matching potentially could be used to estimate camera pose for the planar object.

We further evaluate the quality of the proposed extension to Maximally Stable Extremal Region (MSER) tracking by analyzing the repeatability of the obtained detections. We used a sequence provided with ground truth from [2], which allows a direct comparison of feature and shape based tracking results. Figure 2 compares results over different Δ parameters. As can be seen especially for lower Δ values (including more and less stable regions), the original approach frequently confuses extremal regions which leads to lower repeatability scores.

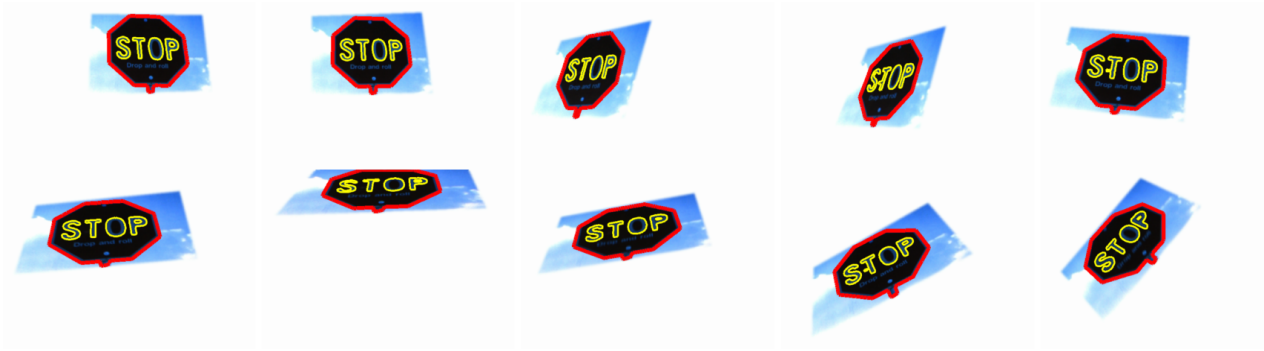

Figure 1. Shape Guided MSER tracking results for a traffic sign. Each of the initialized five MSERs (four MSER+ shown in yellow and one MSER- shown in red) is robustly tracked through the entire sequence yielding 100% repeatability.

4 Conclusions

This paper described an extension for tracking Maximally Stable Extremal Regions (MSERs) through image sequences. We outlined how fast shape matching is used to improve detection repeatability. Therefore, the proposed method is suited for applications like object detection or action recognition which demand robust interest point tracks through image sequences. Furthermore, our proposed method allows tracking of weakly textured, planar objects through sequences, where the implicitly provided point correspondences can be used to estimate the pose of the object. Future work will focus on analyzing the applicability for subsequent vision task as 3D reconstruction or object recognition and localization.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[2] M. Donoser and H. Bischof. Efficient maximally stable extremal region (MSER) tracking. In *Proceeding of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 553–560, 2006.

[3] M. Donoser, H. Riemenschneider, and H. Bischof. Efficient partial shape matching of outer contours. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, 2009.

[4] P. F. Felzenszwalb and J. D. Schwartz. Hierarchical matching of deformable shapes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[5] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 145–151, 2009.

[6] H. Ling and D. W. Jacobs. Using the inner-distance for classification of articulated shapes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 719–726, 2005.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 384–393, 2002.

[8] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 128–142, 2002.

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

[10] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006.

[11] D. Nistér and H. Stewénius. Linear time maximally stable extremal regions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 183–196, 2008.

[12] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003.

[13] T. Tuytelaars. *Local Invariant Features for Registration and Recognition*. PhD thesis, University of Leuven, 2000.

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