Beyond MSER: Maximally Stable Regions using Tree of Shapes

Petra Bosilj[1]
petra.bosilj@irisa.fr
Ewa Kijak[2]
ewa.kijak@irisa.fr
Sébastien Lefèvre[1]
sebastien.lefevre@univ-ubs.fr

[1] Université de Bretagne-Sud
IRISA
Vannes, France
[2] Université de Rennes 1
IRISA
Rennes, France

Detection of local features which are distinctive, invariant and discriminative is used to construct compact image representations in many computer vision applications. Achieving robustness against viewpoint change motivated the development of affine invariant detectors responding to image gradient or contrast changes, edges or corners. We focus on the Maximally Stable Extremal Regions (MSER) detector [3] which responds to blobs of high contrast to produce affine invariant, distinctive arbitrary shaped regions. Exploiting the tree-based MSER computation algorithm [6], we replace the Min and Max-trees [7] in the algorithm with the Tree of Shapes [5], thus changing the pixel ordering used for region extraction.

Min and Max-trees [7] represent the composition of complex images by encoding hierarchical relations of regions on various scales. The leaves of the Min-tree correspond to local image minima, while the inner nodes are (maximal) connected regions $\mathcal{R}_k$ at gray level k, such that all region pixel intensities $f(p)$ are lower than k. The root region $\mathcal{R}_{max}$ at the highest gray level covers the whole image. Distance between two nodes is their gray level difference: $d(\mathcal{R}_k,\mathcal{R}_l) = |l-k|$. The Max-tree is the *dual hierarchy*, corresponding to the Min-tree of an inverted image $-I$.

Extremal regions used in MSER computation [3] correspond to the Min and Max-tree nodes. *Minimal* extremal regions $\mathcal{R}_k$ are connected regions in which all the elements on the outer boundary have strictly greater intensity than all the adjacent region elements, and are contained in the Min-tree. Similarly, the Max-tree comprises the *maximal* extremal regions. MSER computation is based on finding the local minima of the stability function $q(\cdot)$ for the extremal regions along the nested sequences:

$$q(\mathcal{R}_k) = \frac{|\mathcal{R}_{k+\Delta} \setminus \mathcal{R}_k|}{|\mathcal{R}_k|}. \quad (1)$$

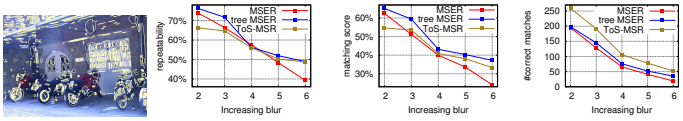
where $|\cdot|$ denotes cardinality. Larger Δ values require region stability through a greater range of gray levels. The region $\mathcal{R}_{k+\Delta}$ is determined from the sequence of nested regions to be the largest region such that $d(\mathcal{R}_k,\mathcal{R}_{k+\Delta}) \leq \Delta$, and found among the ancestral regions of the region $\mathcal{R}_k$ in the corresponding tree. The stability function $q(\cdot)$ can then be calculated concurrently with tree construction [6].

Substituting the Min and Max-tree with the Tree of Shapes (ToS) [5] became viable with the introductions of a near-linear construction algorithm [1]. The ToS models both dark and bright structure by encoding the image composition in terms of shapes and their contrast with their background. It has the *self-dual* property, being unchanged if constructed for the inverted image $-I$. In order to construct a *ToS based Maximally Stable Regions (ToS-MSR)* detector, we have to define the region distance to be used for ToS.

The leaves of ToS correspond to both image maxima and minima, and the regions of the hierarchy are obtained by filling the holes in the extremal regions in Min and Max-trees. A shape $\mathcal{R}$ is the direct parent of the shape $\mathcal{Q}$ if $\mathcal{R}$ is the smallest shape containing $\mathcal{Q}$. Any region $\mathcal{R}$ corresponding to an inner node is composed of the image elements of all of its children and some additional elements, which are always on the same gray level k. The node whose all additional elements are on the level k is referred to as $\mathcal{R}_k$. The distance between any two nodes of ToS in a vertical relation is then defined based on the pair-wise difference between the neighboring node levels. The distance between regions $\mathcal{R}_k \subseteq \mathcal{R}_l$ amounts to the sum of consecutive distances of all the nested regions on a path between those regions, and is equal to:

$$d(\mathcal{R}_k,\mathcal{R}_l) = |k-k_0| + |k_0-k_1| + \cdots + |k_x-l|. \quad (2)$$

The constructed ToS-MSR detector returns slightly more detected regions than MSER, which are still of arbitrary shape but better centralized after affine construction of measurement regions (cf. Fig. 1(a)). We evaluate our detector in the matching framework of Mikolajczyk et al. [4], as well as in a retrieval setup, and compare it to the MSER implementation provided for [4], as well as a tree-based MSER implementation. Typically, we achieve comparable repeatability and matching scores as the MSER detectors with a 20–40% more correct matches (shown in Figs. 1(b)–1(d)). This mitigates one of the main drawbacks of MSER, which occasionally returns too few regions even for applications where a small number of regions is an advantage, such as matching or retrieval. In our retrieval experiments using the VLAD indexing scheme [2], we evaluate the performance in terms of *mean average precision (MAP)* on *INRIA 'holidays'* and *'oxford5k'*. The vocabulary was built using the *'paris6k'* dataset. In addition to the small but consistent increase in the number of features, our ToS-MSER detector outperforms both MSER versions in terms of MAP (shown in Tab. 1). The improvement in the retrieval experiments, output in terms of arbitrary shapes organized in a single hierarchy as well as slightly increasing the number of MSER outputs prompts further investigation of component trees for region detection.



(a) (b) (c) (d)

Figure 1: Performance of ToS-MSR on the *'bikes'* dataset [4] (typical across other framework datasets). Detector responses for image 1 are shown in (a). Repeatability and matching scores are displayed in (b) and (c). The number of correct matches per image pair can be found in (d).

detector	*'holidays'* features	*'holidays'* MAP mean	*'holidays'* MAP high	*'oxford5k'* features	*'oxford5k'* MAP mean	*'oxford5k'* MAP high
MSER	914.78	0.434	0.451	874.02	0.227	**0.252**
tree MSER	1000.57	0.419	0.431	931.08	0.222	0.232
ToS-MSR	1295.85	**0.451**	**0.462**	1160.98	**0.239**	0.250

Table 1: Results of the image retrieval experiments, using *'paris6k'* for vocabulary construction, and *'holidays'* and *'oxford5k'* for validation. Mean and best MAP values for 8 randomly reinitialized vocabularies.

[1] T. Géraud, E. Carlinet, S. Crozet, and L. Najman. A Quasi-linear Algorithm to Compute the Tree of Shapes of nD Images. In *ISMM 2013*, pages 98–110. Springer, 2013.

[2] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR 2010*, pages 3304–3311. IEEE, 2010.

[3] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *BMVC 2002*, pages 384–396, 2002.

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

[5] P. Monasse and F. Guichard. Scale-space from a level lines tree. *J. Visual Commun. and Image Represent.*, 11(2):224–236, 2000.

[6] D. Nistér and H. Stewénius. Linear time maximally stable extremal regions. In *ECCV 2008*, pages 183–196. Springer, 2008.

[7] P. Salembier, A. Oliveras, and L. Garrido. Antiextensive connected operators for image and sequence processing. *IEEE T. Image Process.*, 7(4):555–570, 1998.