Extremal Regions Detection Guided by Maxima of Gradient Magnitude

Mehdi Faraji, Jamshid Shanbehzadeh, Kamal Nasrollahi, and Thomas Baltzer Moeslund

***Abstract*— A problem of computer vision applications is to detect regions of interest under different imaging conditions. The state-of-the-art maximally stable extremal regions (MSERs) detects affine covariant regions by applying all possible thresholds on the input image, and through three main steps including: 1) making a component tree of extremal regions' evolution; 2) obtaining region stability criterion; and 3) cleaning up. The MSER performs very well, but, it does not consider any information about the boundaries of the regions, which are important for detecting repeatable extremal regions. We have shown in this paper that employing prior information about boundaries of regions results in a novel region detector algorithm that not only outperforms MSER, but avoids the MSER's rather complicated steps of enumeration and the cleaning up. To employ the information about the region boundaries, we introduce maxima of gradient magnitudes (MGMs) which are shown to be points that are mostly around the boundaries of the regions. Having found the MGMs, the method obtains a global criterion for each level of the input image which is used to find extremum levels (ELs). The found ELs are then used to detect extremal regions. The proposed algorithm which is called extremal regions of extremum levels (EREL) has been tested on the public benchmark data set of Mikolajczyk. The obtained experimental results show that the inclusion of region boundaries through MGMs, results in a detector that detects regions with high repeatability scores and is more robust against noise compared with MSER.**

***Index Terms*— Maxima of gradient magnitude (MGM), maximally stable extremal region (MSER), extremal regions of extremum levels (EREL), feature detection.**

I. Introduction

MANY computer vision applications such as image registration, object recognition, image retrieval, to name a few, employ a feature extraction phase in order to represent the image by a set of vectors capable of conveying the pertinent information of the image. Feature extraction usually includes two steps of detection and description which are followed by a matching step. Some known works such as bag-of-features [2], multi-resolution bag-of-features [3], and hyperfeatures [4] have substituted the matching phase with two stages, namely feature clustering and constructing the histogram of visual code-word occurrences. The steps involved in feature extraction are highly related to each other, thus, more efficient results from one step lead to more efficient outcomes from the next one. The outset of the process which is the detection phase has an absolutely crucial role in obtaining more stable overall performance. Thus, presenting an efficient detection method, which is exactly the focus of this paper, is of great importance in many computer vision applications.

Recent detectors usually report the locations of the interest points and their surrounding regions by several geometric parameters. These parameters which are obtained by the second moment matrix can approximate the shape of the detected region of interest. Particularly, the idea of representing the elliptical regions has been proposed by Lindeberg [5]–[8] and Gårding and Lindeberg [9]. He used this idea to represent the detected regions by the Laplacian and the scale-space determinant of the Hessian and also by the gray-level blobs. Harris-Affine/Laplace detector [10]–[12], which detects interest points by auto-correlation matrix across scale space, and Hessian-Affine [10], which detects interest point based on the Hessian matrix across scale space, are two notable detectors that have followed the Lindeberg's idea of determining the surrounding region around the interest points. Similarly, Edge-Based Region (EBR) [13] and Intensity-Based Region (IBR) [14] are suitable for affine regions detection. EBR is a model-based approach that tries to find some structures in the image while IBR evaluates a function for all rays emanated from an extremum point to find a region. A comprehensive evaluation of these detectors can be found in [1].

The fundamental theory of scale-space representations, was first thoroughly discussed by Lindeberg [6], [15]. His outstanding works with scale-selection and feature detection [7], shape affine adaptation [16], and the theoretical concepts of image matching [17], has contributed to the literature significantly. In [15], he well identified the underlying concept for what later proposed as MSER [18], by defining the *gray-level blob* as the 3D volume delimited by the gray-level surface and the baselevel [15]. Therefore, based on the mentioned concept, the well-known region detector, Maximally Stable Extremal Regions (MSER) [18], which has been the inspiration for the proposed algorithm in this paper, has been introduced. Given an input image, MSER, which is a level set based algorithm, thresholds the image with all possible threshold values. The goal of MSER is to find a range of thresholds in which the

Manuscript received January 27, 2015; revised June 3, 2015 and August 3, 2015; accepted August 25, 2015. Date of publication September 7, 2015; date of current version October 6, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonggang Shi.

M. Faraji and J. Shanbehzadeh are with the Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran 14911-15719, Iran (e-mail: farajimhd@gmail.com; shanbehzadeh@gmail.com).

K. Nasrollahi and T. B. Moeslund are with the Visual Analysis of People Lab, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark (e-mail: kn@create.aau.dk; tbm@create.aau.dk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The video is a visual abstract of EREL. The total size of the videos is 23 MB. Contact farajimhd@gmail.com for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2477215

regions are more stable than the regions appearing outside the range. A region is considered stable if its area in a level changes slighter than its area belonging to other levels.

MSER has been used in many computer vision applications, like, stereo vision in [18]–[21], object recognition [22]–[24], 3D segmentation [25], lane detection and tracking systems [26], [27], real-time image segmentation [28], finger and hand detection [29], image registration in remote sensing [30], [31], [31], large-scale image retrieval [32], video stabilization [33], and text detection [34]–[36]. Several authors [37]–[39] have proposed improvements on the implementation of MSER. They have concentrated on the algorithmic details of MSER and concurrent construction of the union finding forest and the components' (extremal region) tree. There have also been lots of extensions over the original version of MSER, as those in [40]–[46]. In [43], an integrated algorithm, ED-MSER, is proposed which combines MSER [18] with SIFT [47] and a filtering strategy. The notion of enclosed regions detected by setting several thresholds was introduced in [44]. Enclosed regions consist of External Enclosing Regions (EER) and Internal Enclosed Regions (IER). Recently, MSER has been adapted to work with scale-space theory as for example in [45], [46], and [48]. Forssen and Lowe [48] introduced a multi-resolution version of MSER and used it to construct a descriptor for detected extremal regions. In [45] a multi scale version of MSER called MMSER has been proposed in which the selected regions should not only be stable in their own levels, but also in the pyramid scale space based on a local minimum of stability criterion. In [46], MSER is employed on the Difference of Gaussian (DoG) which have been inspired by the concept of original scale-space and the gray-level blobs proposed by Lindeberg [15]. The stable regions are selected if the barycenter of a region is surrounded by at least ten barycenters of other regions in its adjacent scales. In [41] Maximally Stable Color Regions (MSCR) has been introduced and obtained by an agglomerative clustering of image pixels, which models the distribution of edge magnitudes.

The topology of a manifold can be investigated by looking at the differentiable functions of that manifold. Milnor [49] explained Morse theory using a torus tangent to a plane. The variation of the topology of a segment of the torus as a function of height above the plane can reflect the general topology of the torus. In image processing and computer vision, the image can be regarded as a landscape height map (Lindeberg [15] considered a non-degenerate gray-level function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ at a fixed level of scale). For each height value, a set of connected pixels (called extremal regions) can be extracted. The set of all extremal regions belong to each height, constitutes a finite set of extremal regions. We propose an innovative method, to select a fair number of extremal regions from the finite set based on a global estimation ratio between boundaries variation and surface variation.

As illustrated in [1], MSER obtained low repeatability when blur transformations happen which can be mostly because of its stability criterion. The stability criterion considers mainly the area of the regions as the crucial parameter. So, the stable regions are those regions with the least change during a range of thresholds. This idea considers no information about the boundaries of the regions during the detection of stable regions. Particularly, blur transformation manipulate the boundaries of the regions, so the area-based stability criterion of MSER can result in low repeatability. To deal with this problem Kimmel *et al.* [50] proposed another stability criterion based on the length of the boundaries. It can be seen that various types of transformations may need different kind of stability criterion to perform well. Another issue with MSER is that it mostly include a rather complicated step of enumeration followed by a cleaning up step. The proposed system in this paper, is an improved version of [51] (Fig. 1) that presents a *completely novel region detection algorithm* in which the above-mentioned issues are tackled by including the information of the region boundaries. To do so, we have introduced a kind of interest points, *Maxima of Gradient Magnitudes* (MGMs), Fig. 2, which are mostly concentrated around the edges (region boundaries). Therefore, they are used as prior information for detecting invariant regions in our region detection algorithm. The experimental results on a challenging benchmark dataset show superiority of the proposed algorithm over the state-of-the-art region detection algorithms.

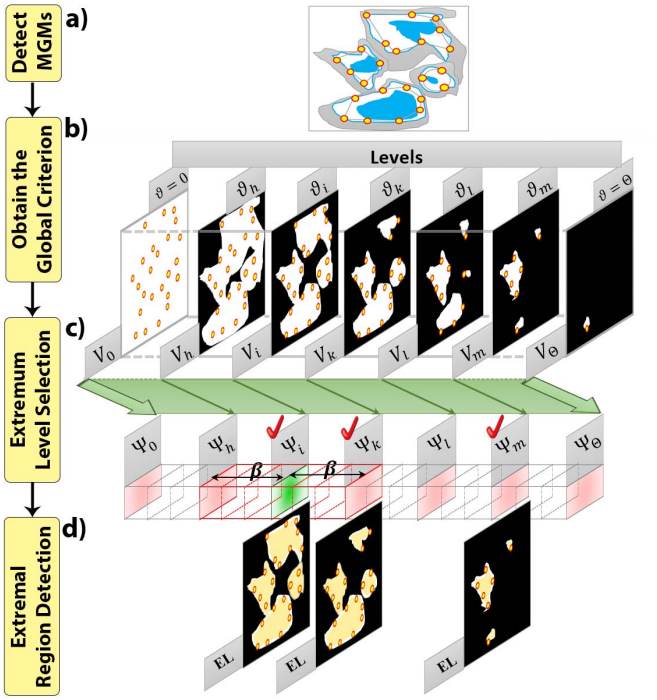

Fig. 1. The block diagram of the proposed system. a) Provides information about edges of the image by detecting MGMs from the input image. b) Calculates a global stability value for each level of the image and stores the values in vector Ψ. c) Selects each level that has a local maximum Ψ. d) Finds the connected components of the selected levels which include at least one MGMs.

The rest of the paper is as follows. The details of the proposed algorithm and the introduced MGM points are explained in the next section. The experimental results are given in Section III. The time complexity analysis of the proposed method is presented in section IV. The effect of the parameters of the method and the advantages of the proposed

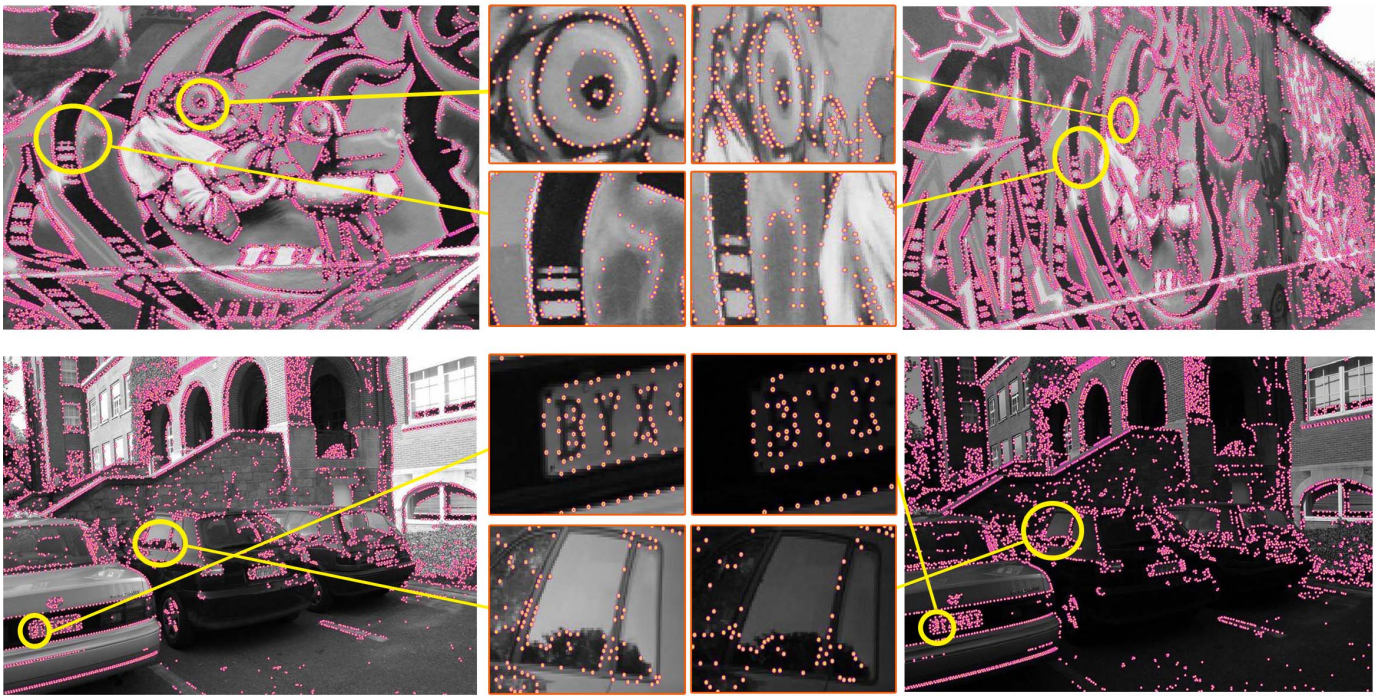

Fig. 2. MGMs for two different parts of the "graf" and the "leuven" images of [1].

method over MSER is discussed in section V. Finally the paper is concluded in Section VI.

II. The Proposed Method

The block diagram of our novel proposed algorithm is shown in Fig. 1. We first detect MGM points (Fig. 1(a)). Then, we binarize the image by applying all possible thresholds on it and obtain a Global Criterion (GC) based on the information provided by both white pixels of the thresholded image and the MGMs. The obtained thresholds intersected with the detected MGMs are shown in Fig. 1(b). After that, we select a number of Extremum Levels (ELs) based on the GC vector. The selection process which finds local extrema of the GC is shown in Fig. 1(c). In the last step (Fig. 1(d)) we detect extremal regions in only those selected ELs that intersect with at least one MGM. The above steps are explained in the following subsections, respectively.

A. Maxima of Gradient Magnitude (MGM)

Given an input image, I, we first obtain its gradient, ∇I, by a simple gradient filter like Soble. An MGM is a point $p(x, y)$ in ∇I that has two conditions: first, it has a maximum value of the gradient magnitudes among their local neighborhood points with radius r, as:

$$\|\nabla I(p)\| > \|\nabla I(P)\|, \quad \forall P \in N(p, r) \tag{1}$$

where $N(p, r) = \{P | \forall P \in I, \|P - p\| < r\}$ is the neighbor function. Second, the mean of the gradient magnitudes of its neighbor points, should be larger than a threshold:

$$E[\|\nabla I(N(p, r))\|] \geq \alpha \cdot \tau \tag{2}$$

where $E()$ is a mean function, and α is an arbitrary coefficient which controls the strength of the of resulted points. Here, τ is a suitable threshold value that used for thresholding the gradient magnitudes image. To find the value of τ we use the *isodata* method [52]. First, we assume that the mean of the image gradient magnitude is an initial point in its histogram that separates foreground from the background. Then, the mean of each distribution is calculated and the average of both means becomes the new threshold. This process continues iteratively (only takes a few iterations) until the threshold value is not changed. Therefore, such a threshold is achieved by a simple iterative algorithm that runs in constant time. In addition, it can keep a fair amount of high informative points along edges. As it can be seen, the process of finding τ is a global function and works on the histogram of the gradient magnitudes of the image. To have more robust results, we first equalize the histogram of the input image prior to the thresholding.

Checking the above two conditions for all the pixels of the input image, a binary image, M, can be generated in which the positions of the MGMs are highlighted. So, we identify that a point termed $p(x, y)$ is an MGM if $M(x, y) = 1$, where, x and y indicate the location of the point.

Fig. 2 shows two example images and the MGMs extracted from them. The high repeatability of MGMs can be seen in zoomed boxes of the Fig. 2. Even for changes in strength of the light source (the second row of Fig. 2), it is clear that MGMs can be easily detected. Therefore, although we employ MGMs as prior information about the boundaries here, one can use them as pure *interest points* in a specific application.

The resulted $M(x, y)$ image from the above process is then used in the next step of the proposed system to obtain a stability criterion.

B. Global Criterion (GC)

Following the block diagram of the proposed system in Fig. 1, having found the MGMs, the next step is obtaining a GC. Two sets of regions, Q^- and Q^+, can be detected from an input image when any type of level set methods, like the one used in this paper, is employed:

- The first set, Q^-, contains regions that evolve from brighter surfaces to darker boundaries. These regions can be detected from the original image by thresholding at different levels. Each of these thresholds result in a binary image, $T_\vartheta^-(x,y)$, as:

$$T_\vartheta^-(x,y) = \begin{cases} 1 & \text{if } I(x,y) \geq \vartheta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

  where ϑ is the current threshold. The range of the thresholds depends on the number of the bits used per pixel.
- The second set, Q^+, contains those regions that evolve from darker surfaces to brighter boundaries. To detect these regions we use:

$$T_\vartheta^+(x,y) = \begin{cases} 1 & \text{if } I(x,y) \leq \vartheta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We should calculate a GC vector for both types of $^+$ and $^-$ regions. Since the steps required are independent of the type of the regions, we only explain them for the $^+$ type of regions. The reader can do the same process for the $^-$ type of regions.

We intend to have a GC with a more distinguishable capability to specify the levels that have the most stable edge variations in proportion to their adjacent levels. However, we should also consider that having only a lot of edge points in a level cannot precisely determine a suitable invariant level to extract covariant regions. If there are small variations in the area of the regions, the edge points variations that we are representing by MGMs variations, cannot discriminate a global invariant level. So, to find levels with most stable edge variations, we need to monitor the the concurrent changes of:

- the total number of the newly appeared white pixels in the thresholded image
- the obtained number of the MGM points in each level that intersect with available pixels in that level.

In fact, the former is an interpretation of the histogram of the image, and the latter is the histogram of the MGM points. So, we define the ratio between these two factors in a V^+ function, defined as:

$$V_\vartheta^+ = \frac{h_\vartheta(I(x,y)\cdot M(x,y))}{\epsilon + h_\vartheta(I(x,y))}, \quad 0 \leq \vartheta \leq \Theta \tag{5}$$

where the $h_\vartheta(I(x,y)\cdot M(x,y))$ and $h_\vartheta(I(x,y))$ represent the histogram of the MGMs and the histogram of the image at level ϑ, respectively. To avoid devision by zero, an ϵ value is used. The parameter Θ indicates the maximum possible intensity value, i.e., for an 8-bit image, $\Theta = 2^8 - 1$.

The underlying variations of V^+ are suitable clues for indicating levels in which the variation of the edge points over the surface area become stable. Therefore, we employ the same concept of the local stability of MSER [18], but in a global consideration. Using vector V^+, the GC$^+$, Ψ^+, is achieved:

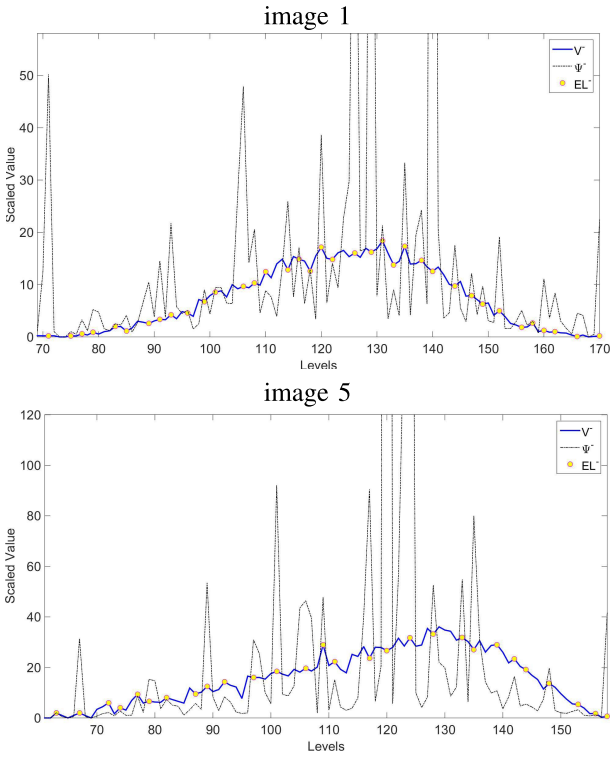

Fig. 3. Selected ELs for two different blur transformations of "trees" image set of [1]. Note that $\beta = 1$ and the values of both vectors have been scaled.

$$\Psi_\vartheta^+ = \frac{V_\vartheta^+}{|V_{\vartheta+1}^+ - V_{\vartheta-1}^+|} \tag{6}$$

Similarly, the same process can be performed on the other type of regions ($^-$) to obtain the Ψ^-. However, because the second phase is performed on the inversed of the image, instead of using h_ϑ, we use the reverse of vector h_ϑ, which is $h_{[\Theta-\vartheta]}$.

C. Extremum Levels (ELs) Selection

Following the block diagram of the algorithm in Fig. 1, having obtained the Ψ_ϑ^+, we need to find ELs$^+$. A level like ϑ belongs to the set of EL$^+$ if its Ψ_ϑ^+, is a local maximum. To select EL$^+$, each cell of Ψ_ϑ^+ is hence compared with its β previous cells and β subsequent cells. β shows the radius of the neighborhood window and represents the number of adjacent levels which are involved in the process of local maxima selection (Fig. 1). For all of our experiments throughout the paper, we set $\beta = 1$. The calculated GC^- vector and the selected ELs have illustrated in Fig. 3. It can be seen that selected ELs are located in the stable trends of the blue graph, which is V^-.

D. Extremal Regions Detection

Extremal regions can be finally detected from each elements of EL$^+$ and EL$^-$ by any arbitrary connected component analysis algorithms or a labeling strategy. Note that for each vector (EL$^+$ and EL$^-$), we run the algorithm separately.

So, the following explanation should also be considered for EL^-. Having a high number of detected extremal regions implies the importance of applying a filter on them to efficiently select more repeatable regions. To do so, the algorithm starts from the first indicated EL^+, extracts the extremal regions of that level and chooses only those extremal regions that intersect with at least one MGM. After that, the MGMs belonging to the selected extremal regions will be ignored in the next ELs^+. This process continues till all elements of ELs^+ be processed. It should be noted that the number of elements in ELs^+ are related to the radius of the neighboring window. For instance for $\beta = 5$ it yields less than 20 elements. If a region intersect with no MGM, it shows that either the region has been detected wrongly because of the presence of the noise, or the region is not stable enough. So, MGMs actually help both selecting regions and cleaning up the unwanted extremal regions. However, it should be noted that no direct cleaning up nor clustering and enumeration is performed by EREL. Since, our proposed method detects extremal regions belonging to both ELs^+ and ELs^-, we call it: Extremal Regions of Extremum Levels (EREL).

III. Experimental Results

The experiments reported in this paper have been conducted on the image sequences of the public benchmark dataset of Mikolajczyk, affine covariant dataset [1]. The images in this dataset have gone through different degradation factors, including:

- blur (by "tress" and "bike" sequences) which have been acquired by changing the camera focus
- viewpoint (by "graf" and "wall" sequences) which have been transformed by changing the camera view from a front-to-parallel view to one with significant foreshortening at almost 60 degrees to the camera [1]
- scaling and rotation (by "boat" sequence) which has been acquired by varying the camera zoom at a factor of about 4, and
- JPEG compression (by "ubc" sequence) which is generated using a standard xv image browser with the image quality parameter varying from 40% to 2% [1].

The above dataset has been used to draw comparisons between the proposed method and the competing state-of-the-art method of MSER [18]. To do so, three tests are conducted. In all tests in this paper, we set all parameters of both methods (MSER and EREL) equally, i.e minimum area of a region equal to 30, maximum area of a region equal to $0.01 \times N$ [18], and for the ellipse fitted to the region, as it has been suggested in [1], we double the scale of the fitted ellipse. In addition, the value of the parameter of the method, $\alpha = 1.2$, has been kept unchanged during all of the tests in this section.

- In the first test, we evaluate the performance of the detector and compare it against MSER. The evaluation is based on two common criteria: *repeatability* and *the accuracy of localization* [1].
- In the second test, we compare our detector against MSER, based on the performance of their extracted descriptors. Most of the applications extract descriptors from detected regions prior to any further processing. So, evaluating the descriptor can generally show the performance of the method in real applications. The two main factors for this evaluations are *recall* and *1-precision*.
- In the third test, we compare our system against MSER in the presence of noise.

The evaluation factors in each of the above three mentioned tests and the obtained results are explained in the following subsections.

A. The First Test (Repeatability and Accuracy)

In this test two evaluation factors of repeatability test and overlap error (accuracy of localization) are considered.

1) Repeatability Test: One of the important factors of assessing systems like the one proposed in this paper, is the repeatability criterion of Schmid [53]. It defines how a detector can repeatedly detect the same regions in different images of the same scene (when images are transformed by different geometric and photometric imaging parameters, including those in the used database of [1]). The repeatability score provides a quantity value of the performance including the accuracy of localization and region estimation and is defined as the ratio of the number of region-to-region correspondences and the smaller number of regions detected in one of the images [1]. To locate the corresponding points in planar image pairs a homography matrix is employed which finds the corresponding points in relative locations with an error less than 1.5 pixel. Higher repeatability score and larger number of correspondences indicate a better detector performance. Therefore, reaching a 100% repeatability and representing the more horizontal line by the plot of the repeatability and the number of correspondences are two ultimate goals of an optimal detector. To have a better understanding of the performance of the EREL, its repeatability scores for six image sets of the dataset are shown in Fig. 4 and Fig. 5 for overlap errors of 10% and 40%, respectively. We illustrate the results of the repeatability scores for 10% overlap error in order to show the accuracy of the regions. Not to mention that, reporting the results for 40% overlap error is common in the literature.

It can be seen from Fig. 4 and Fig. 5 that in most of the cases the proposed system outperforms MSER. However, the degree of the improvement changes from one image to another, since the contents of the images in the dataset are very different. For example, the "trees" and the "wall" images represent textured type of scene and contain huge amount of variations. Although, this means that the MGMs can be better detected on the boundaries of the shapes, the concept of detecting blob-like regions is in contrast with this kind of images. Lindeberg [15] has stated this very clearly by saying that the presence of another nearby blob may neutralize a blob or reduce its size which at the end results in decreasing the repeatability of every detector that designed to work based on this theory.

2) Overlap Error: Another important factor in assessing the performance of detectors is the overlap error which shows the

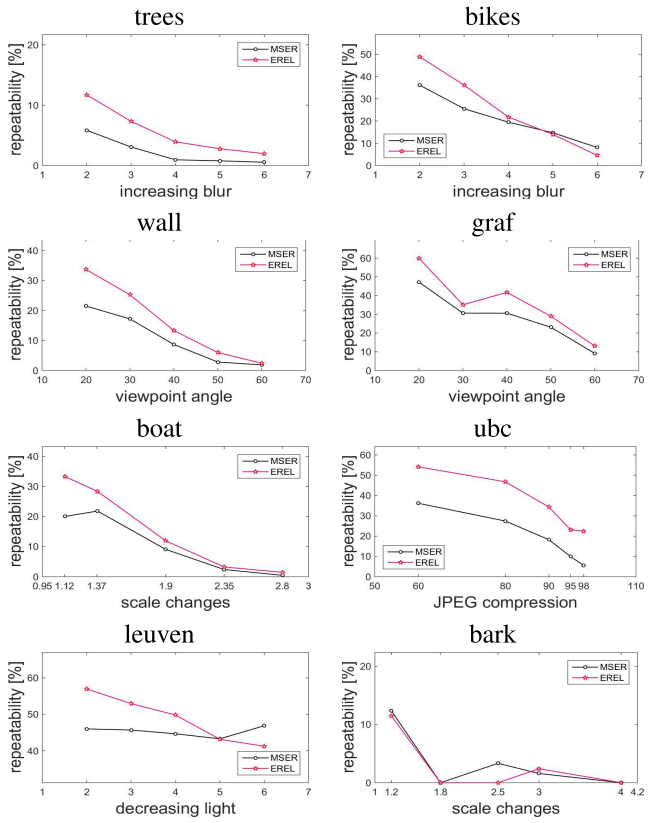

Fig. 4. Repeatability scores achieved by 10% overlap errors for all image sets of dataset [1].

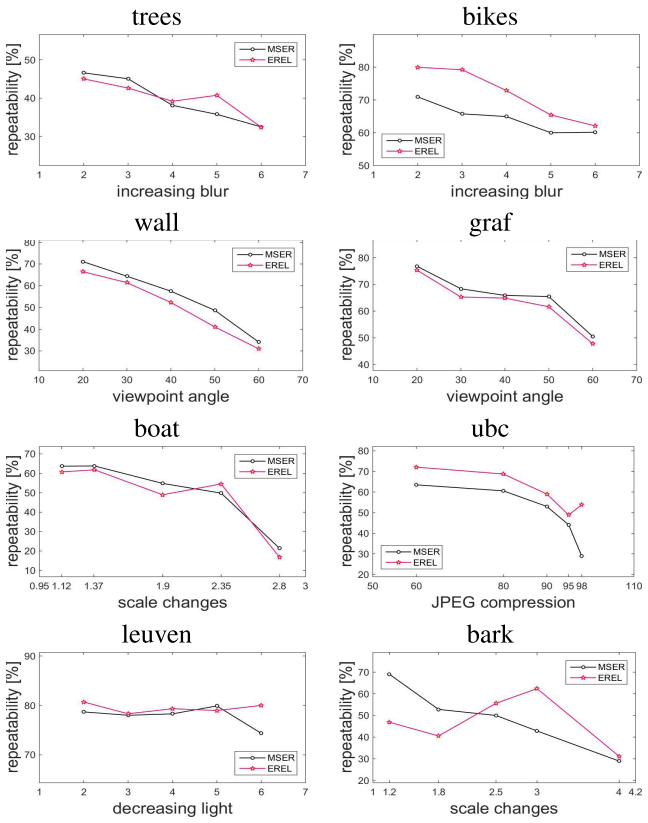

Fig. 5. Repeatability scores achieved by 40% overlap errors for all image sets of dataset [1].

accuracy of the localization and the region estimation. Overlap error, ϵ_o, which is used to find two corresponding regions, if ϵ_o is small enough, is defined:

$$\epsilon_o = 1 - \frac{|R_{\mu_a} \cap R_{(H^T \mu_b H)}|}{|R_{\mu_a} \cup R_{(H^T \mu_b H)}|} \tag{7}$$

where $|.|$ is a function that calculates the area of the region, H is a homography matrix (used to find a point to point correspondences), R_μ is an elliptic region, and $\cap$ and $\cup$ represent the union and the intersection between the regions, respectively. Obviously, smaller overlap error implies greater similarity of the detected regions of the reference image and their counterparts in the transformed images. We have also compared the EREL against MSER by the repeatability score as a function of the overlap error in several image pairs of the Mikolajczyk's dataset of [1]. The results (shown in Fig. 6) reveal that the detected regions by the proposed system in most pairs are the most accurate ones, and it achieves the best repeatability scores.

3) Number of Detected Regions: The repeatability score's superiority of the proposed system to MSER is more valuable when we compare the number of the detected regions of the two methods, because the number of the detected regions can affect the repeatability scores. If a detector detects only few regions one can expect a high repeatability score. On the other hand, detecting a huge number of regions can causes lots of accidentally matched regions [1]. It is shown in Fig. 13(d) that the number of the regions detected by our method is very close to those found by MSER Fig. 13(c), which means that the number of the detected regions by the proposed system is comparable with MSER.

4) Region Size: As it has been demonstrated in [1], having a lot of big regions affect the repeatability scores since they contain more information, they can be matched more easily and consequently, the score will be increased. However, not to mention that the rate of the increase is not very substantial. A method with a poor design can never achieve high performance based on having bigger area sizes. One way to show its effects on the performance or on the repeatability, is to obtain the relative size of the detected regions. Accordingly, we illustrate the relative areas of the detected regions by EREL and compare it with MSER in Fig. 7. It can be seen from this demonstration that the average of EREL's region size are close to MSER in most cases.

B. The Second Test (Descriptors)

The proposed EREL algorithm not only detects regions with high repeatability, but improves the performance of real world-applications. To show this, we extract SURF descriptors [55] from the regions detected by both MSER and EREL. As both methods are able to report all pixels that contribute to a region, we extract SURF descriptors based on the detected extremal regions' pixels (Distinguished Regions [18]). We then compare the performance of the features obtained from these descriptors against each other. The performance is evaluated based on criteria explained in [56], namely *recall* and *precision*.

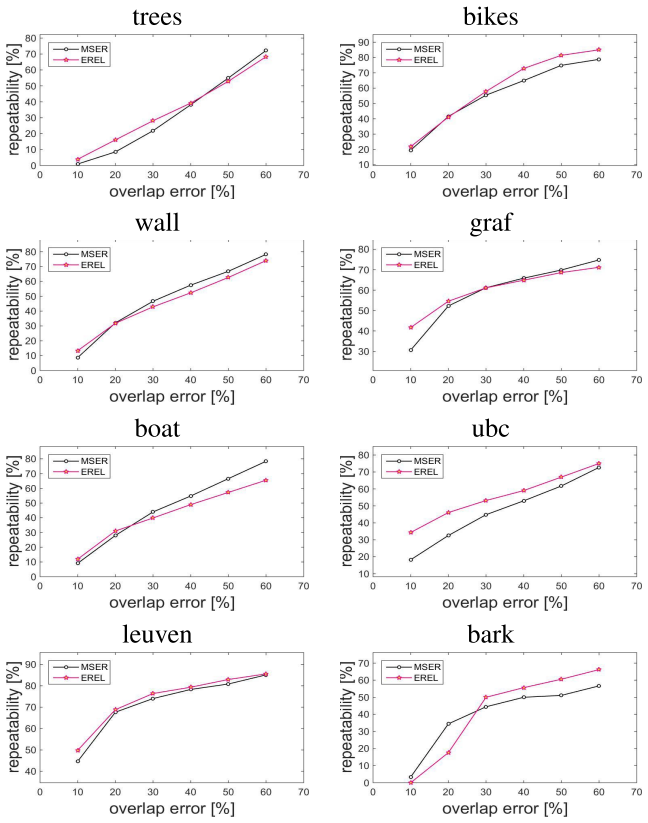

Fig. 6. Repeatability score as function of overlap error for image pairs of the dataset of [1]. All pairs are similar, i.e. $< image1, image4 >$.

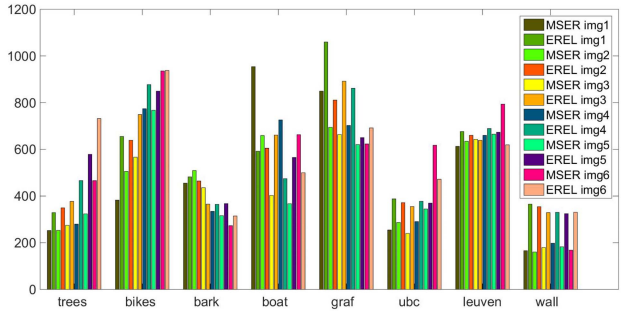

Fig. 7. Comparing the relative area of the regions detected by EREL against MSER [18] for all of the image sets of [1].

The process of descriptor evaluation begins with matching descriptors of the reference image and the transformed image. The matching can be done using different strategies, e.g., 1) Threshold based matching, 2) Nearest neighbor based matching, and 3) Nearest neighbor distance ratio matching (see [56] for further explanations). Although all three strategies can correctly evaluate the performance of descriptors, especially when no specific application is desired, we use the *nearest neighbor based matching* in the matching step due to its slightly better performance [56].

The second step is to determine which of the matched regions by the matching strategy actually are correct matches and which of them are not. This can be done by employing a homography matrix. The result of this step are the number of correct matches and the number of correspondences. These numbers are used to calculate the two important criteria for evaluating the performance of descriptors:

$$recall = \frac{\mathcal{M}_c}{C} \tag{8}$$

$$1 - precision = \frac{\mathcal{M}_f}{\mathcal{M}_c + \mathcal{M}_f} \tag{9}$$

where C represents the total number of correspondences between two images. Here, $\mathcal{M}_c$ and $\mathcal{M}_f$ indicate the number of correct matches and the number of false matches, respectively. Drawing a graph with *recall* on vertical axis and *1-percision* on the horizontal axis (as has been used in [56]) clarifies the performance of methods based on these criteria. These are illustrated in Fig. 8.

The ideal value for *recall* is one. So, higher value of recall in based on the precision implies the superior performance of a method. Considering these facts, it can be seen from Fig. 8 that the performance of the EREL is better than MSER. These results confirm the accuracy of EREL as well. The accuracy of EREL (higher repeatability in lower overlap error), have influenced the descriptor performance and has improved it, i.e., in higher precisions (horizontal values closer to zero), greater recalls are achieved.

Another interesting conclusion can be drawn by comparing the repeatability results (see Fig. 4 and Fig. 5) of "wall", "trees", and "bark" image sequences with the results of its descriptor performance (see Fig. 8). Although the repeatability of EREL is lower than or equal to MSER in some cases, the performance of the extracted descriptor is higher than MSER. This shows that the regions of EREL have been distinctively detected and can be finely matched using a standard descriptor.

C. The Third Test (Presence of Noise)

The presence of noise is an inevitable obstacle of real-world applications. In the following section, we report our evaluation results of EREL against noise contamination and show that our proposed method is more robust to noise when compared to MSER. To do that, we first contaminate all the images of the dataset [1] using Gaussian noise with three different variances ($\sigma^2 = 0.01$, 0.03, and 0.05). Then, we apply MSER and EREL to these noisy images and calculate the repeatability scores from the extracted noisy regions. The obtained scores are illustrated in Fig. 9. It can be seen that, in most occasions, repeatability scores of MSER are lower than EREL when the images are contaminated by noise. However, the performances of both methods drop when the noise variance is increased, which is a natural consequence of noise contamination.

Another comparison is drawn based on the performance of the extracted descriptor (see section III-B) of the noise contaminated EREL and MSER. The results are shown in Fig. 10. It can be seen that recall of EREL in most cases, are higher than MSER for similar noise variances. So, more distinctive descriptors can be extracted from EREL rather than MSER when the input images are noisy which confirms the robustness of EREL in real applications.

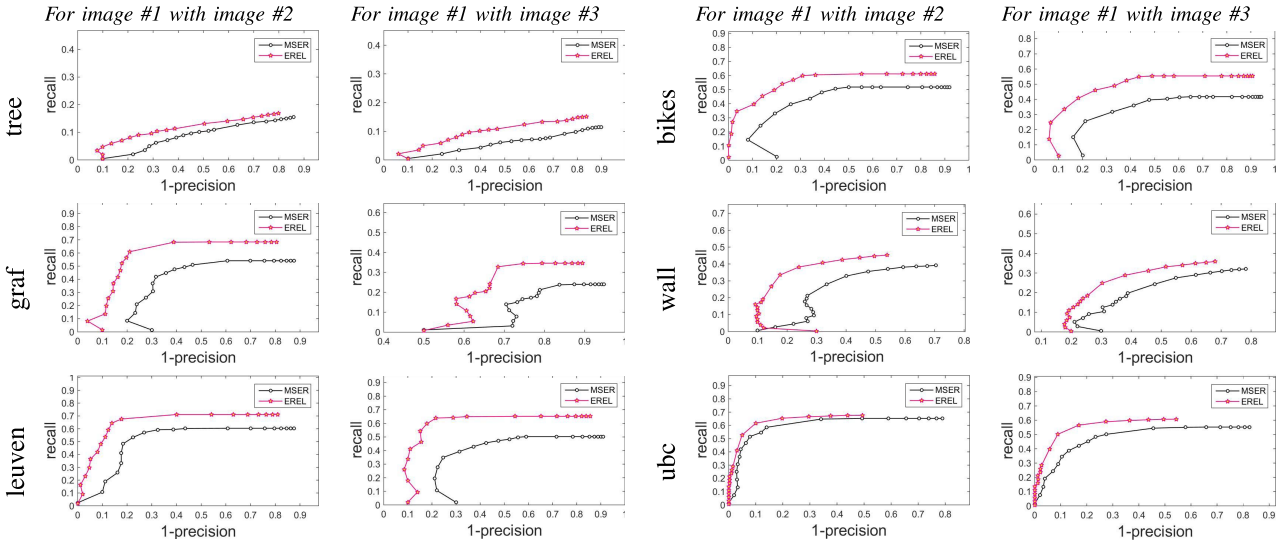

Fig. 8. The performance evaluation of the EREL and MSER based on an extracted SURF descriptor for each of them. Features have been extracted from images of [54] dataset. Each row belongs to a same image sequence. Columns represent the image pairs.

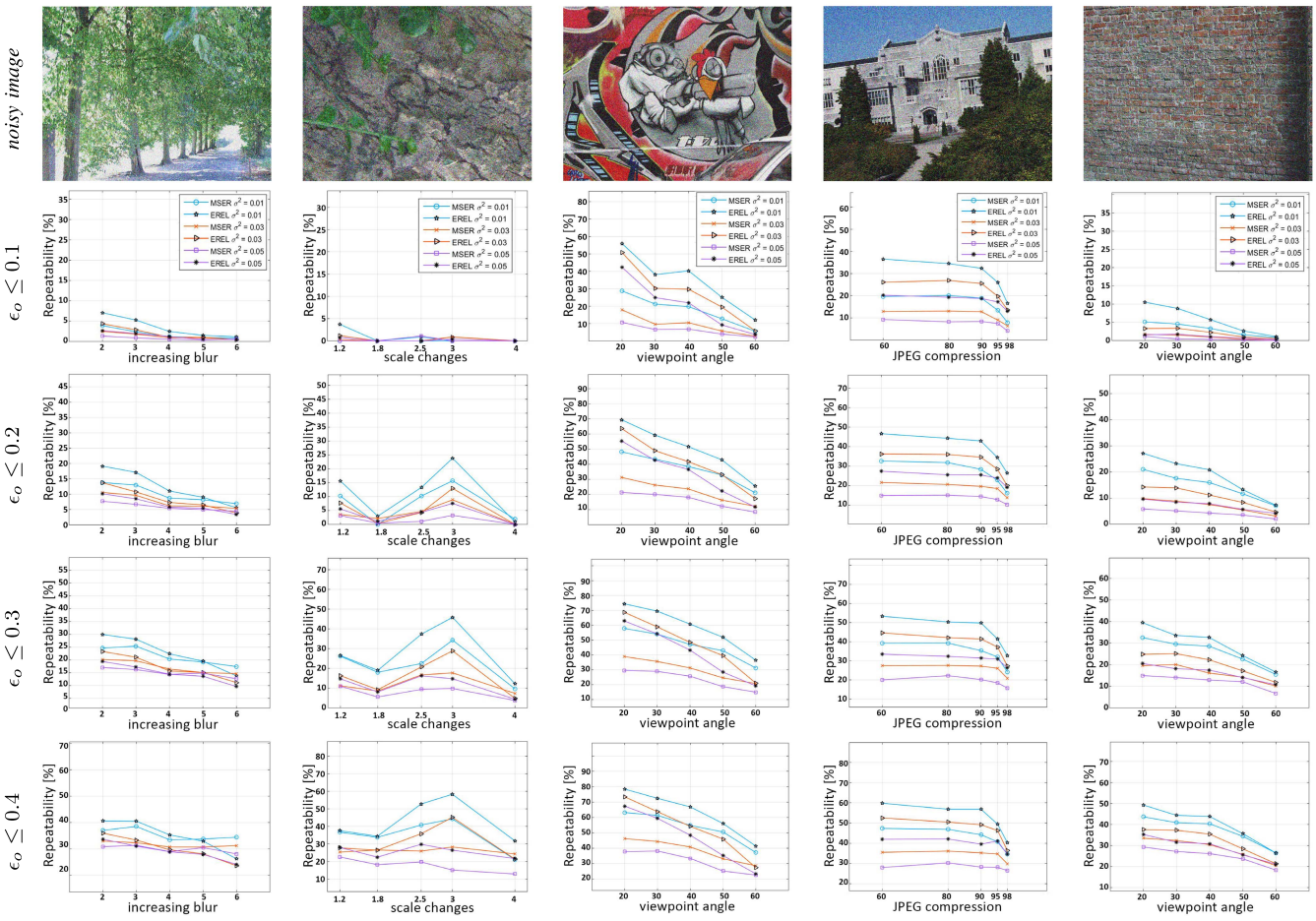

Fig. 9. The repeatability scores of EREL and MSER extracted from noisy images from [1] dataset. Each row represents the results for different overlap errors. Columns indicate the input images.

D. EREL Against Recent Versions of MSER

In this section we further compare the repeatability score of the proposed EREL method against the very recent extensions of MSER which have reported their results on the benchmark dataset of [1]. These systems are TBMR [57] and Enclosed region [44]. TBMR uses a topological approach for region

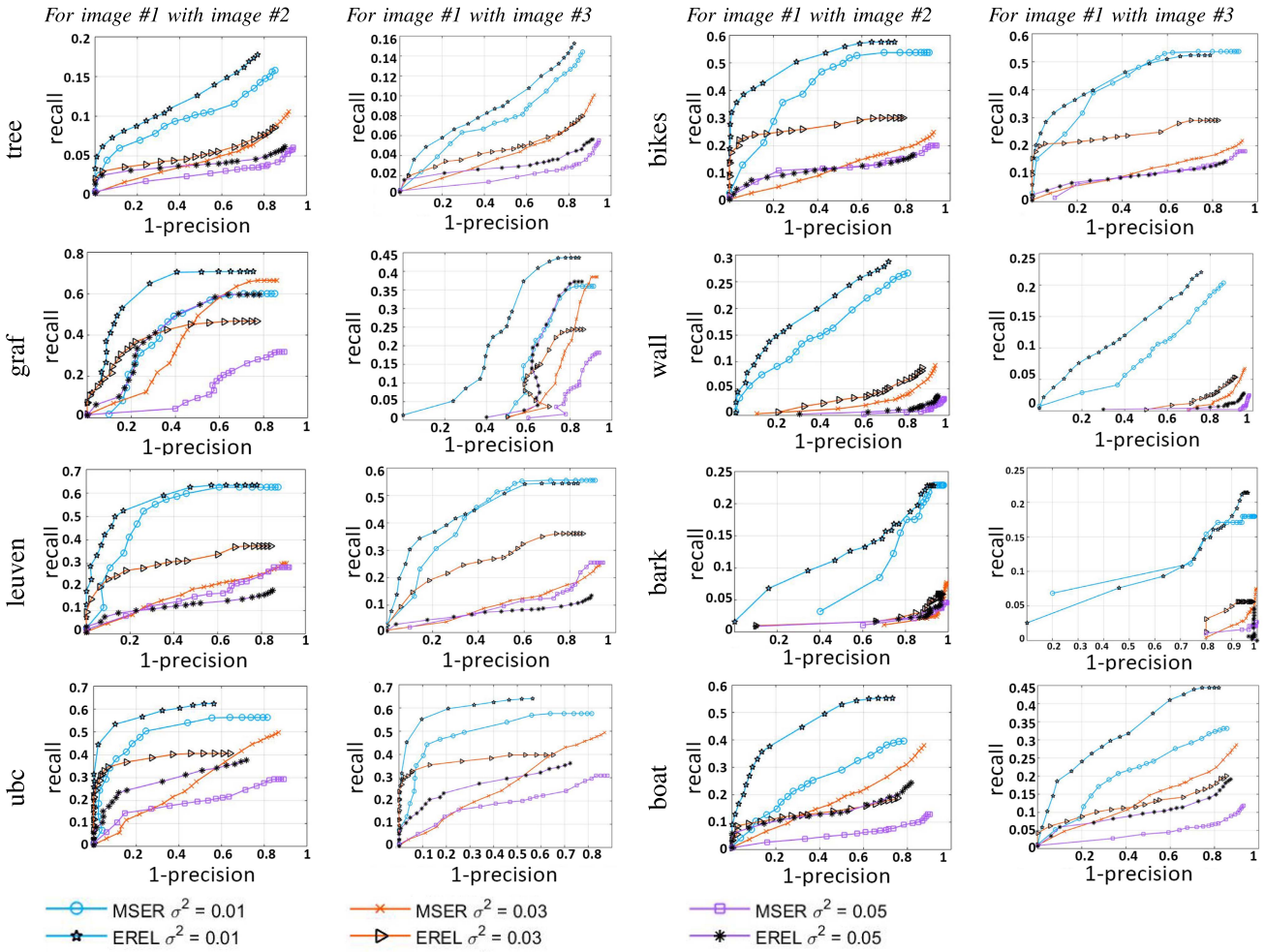

Fig. 10. The performance evaluation of the EREL and MSER based on an extracted SURF descriptor against noise contamination. Features have been extracted from images of [54] dataset. Rows belong to image sequence. Columns represent the image pairs.

detection in which regions are contrasted [57] while Enclosed region based system assumes that an object is enclosed by the same region before and after the degradation and/or transformation [44]. The results of the comparison (for an overlap error of 40%) are shown in Fig. 11. It can be seen from this figure that for structure type scene ("bark") and a blur degraded image ("bikes") EREL has high performance which is close to the other methods.

IV. Time Complexity Analysis

Prior to discuss about the time complexity, we review the steps of our proposed method considering an efficient implementation. Although EREL can be simply implemented based on a parallel methodology, we have implemented it sequentially in order show that it has a linear running time, and also to draw a fair comparison to the other implantations of MSER. In short, EREL detection has four steps including detection of MGM points, obtain two global criteria vectors, extremum level selection, and extremal region detection. EREL can also implemented efficiently based on other ideas, for example using connected component algorithms or using a flood-fill method, however our implementation for each step is as following. Please note that by N we mean the number of pixels in the image. In addition, we neglect every histogram traversal and consider its running time as constant because the number of levels for an 8-bit image is 256 which is too small and takes a few CPU clocks to run.

- *Image Initialization:* At first, we pass every pixels of the image to prepare the histogram of the image and subsequently equalizing it ($O(N)$), then we apply a Sobel filter on the image in order to have a gradient magnitude image ($O(N)$), and finally we construct an integral gradient image [55], [58] ($O(N)$). The overall complexity for this step is $O(N)$.
- *MGMs Detection:* To efficiently extract MGM points from the image, we use the concept of integral image which has been introduced in [58]. As MGMs are extracted from the obtained gradient magnitudes, we use the exploited integral gradient image from the initialization step to calculate the neighborhood average of each pixel in a constant time. So, only four array references happen for each averaging which result in linear time extraction of the MGM points. Finally, during the process of MGM detection, as soon as an MGM found, the histogram of MGMs is updated.

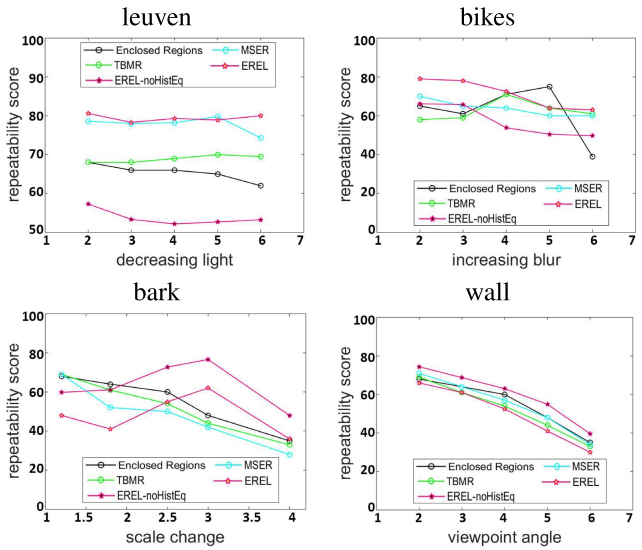

Fig. 11. Comparing the Repeatability scores of EREL with MSER and two very recent methods of TBMR [57] and Enclosed region [44] (Overlap error is 40%). Please note that some of the plots are overlapping. The plots of EREL are obtained based on having a histogram equalization and without a histogram equalization.

- *Obtaining Two Global Criteria:* For calculating the global criteria we need histogram of the image and histogram of the MGMs, see Eq. (6), which have been calculated from the previous step. Therefore, this step takes constant time to compute the vector Ψ^+ and vector Ψ^-.
- *Extremum Level Selection:* Similar to the previous step, local maxima selection from the vector Ψ^+ and vector Ψ^- need a constant time.
- *Extremal Region Detection:* The main part of the implantation of the EREL and its time complexity, depends on this step. An efficient algorithm can significantly decrease the running time of the whole method. So, we use a weighted quick union-find structure with path compression [59] in order to join the connected pixels and sequentially create a tree of image pixels. Reader can refer to [59] for thorough discussion about union-finding. To do so, we first sort pixels in decreasing order. Since we have calculated the histogram of the image, we employ the histogram to sort pixels in a linear time by BINSORT [59], the idea of this type of sorting is similar to [18], [38]. Note that, we just sort the pixel once and for doing union-find on the inversed of the image, we traverse the pixels from N to 1. The complexity of the weighted quick union-find structure is $O(N)$ [59], because every union and find step take constant time and all pixels in the image are checked based on their 4-connectivity. Compared to Nister version of MSER [39], our method is even faster, since it stops doing union as soon as it passes the last specified extremum level (obtained from the previous step of EREL). So, we do not construct the whole forest.

All in all, although the method should run twice (for Q^+ and Q^- regions), the complexity of EREL is $O(N)$. The complexity of the original MSER is $O(NlglgN)$ [18]. Additionally, Nistér and Stewénius [39] proposed a faster

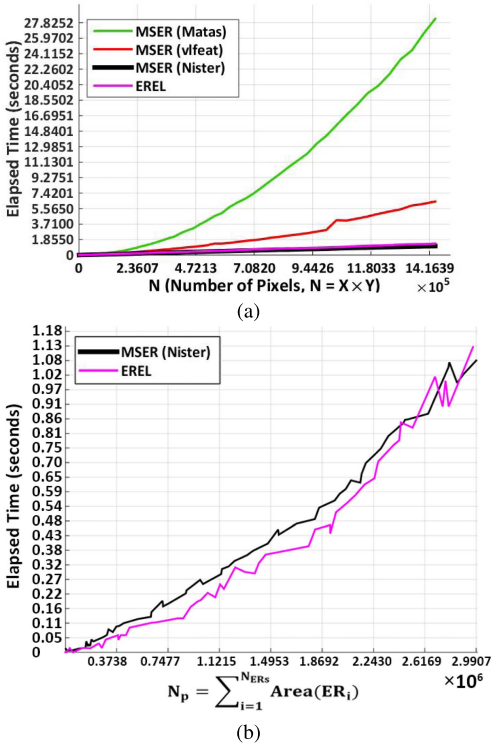

Fig. 12. Comparison of the Running time of the proposed EREL with several implementations of MSER (Matas *et al.* [18], Vedaldi [60], Nistér and Stewénius [39]). (a) The running time based on the increasing number of pixels (N) in the input image. (b) The running time based on the increasing number of detected pixels (N_p) from the input image. Both (a) and (b) belong to a same detection sequence.

algorithm based on another immersion strategy which is almost linear. Another efficient implementation has been proposed in [60]. We compare the actual running time of the EREL with aforementioned three efficient implementations of MSER [18], [39], [60]. The test has been implemented on a laptop with a Core i7 CPU and 16GB of RAM. The results can be seen in Fig. 12. We prepare a test that measures the running time of EREL and three different implementations of MSER on a similar condition but based on two different considerations. The first one is the elapsed time against the number of pixels in the image, N. As we demonstrate in Fig. 12(a), the elapsed time of EREL is almost equal to the MSER proposed by Nister and Stewnius [39] based on N which is linear. Secondly, we consider a criterion that mostly ignored in the literature, i.e. the total number of extracted pixels by the detector. We know that the original MSER [18] and also vlfeat-MSER [60] do not report exact coordinates of the pixels in the region and only provide us with the parameters of the fitted ellipse on that region plus a threshold. In order to get the exact coordinates of the pixels belonging the region, one should perform a flood-fill algorithm to retrieve the pixels of that region, which is a costly process. That is why the two methods in Fig. 12(a) take longer time to be finished. A possible solution for the mentioned issue has been discussed in [37]. They proposed to construct and save the whole tree in a structure called N-Tree Disjoint Set Forest (NDS). This idea again needs to pass each of the detected pixels at

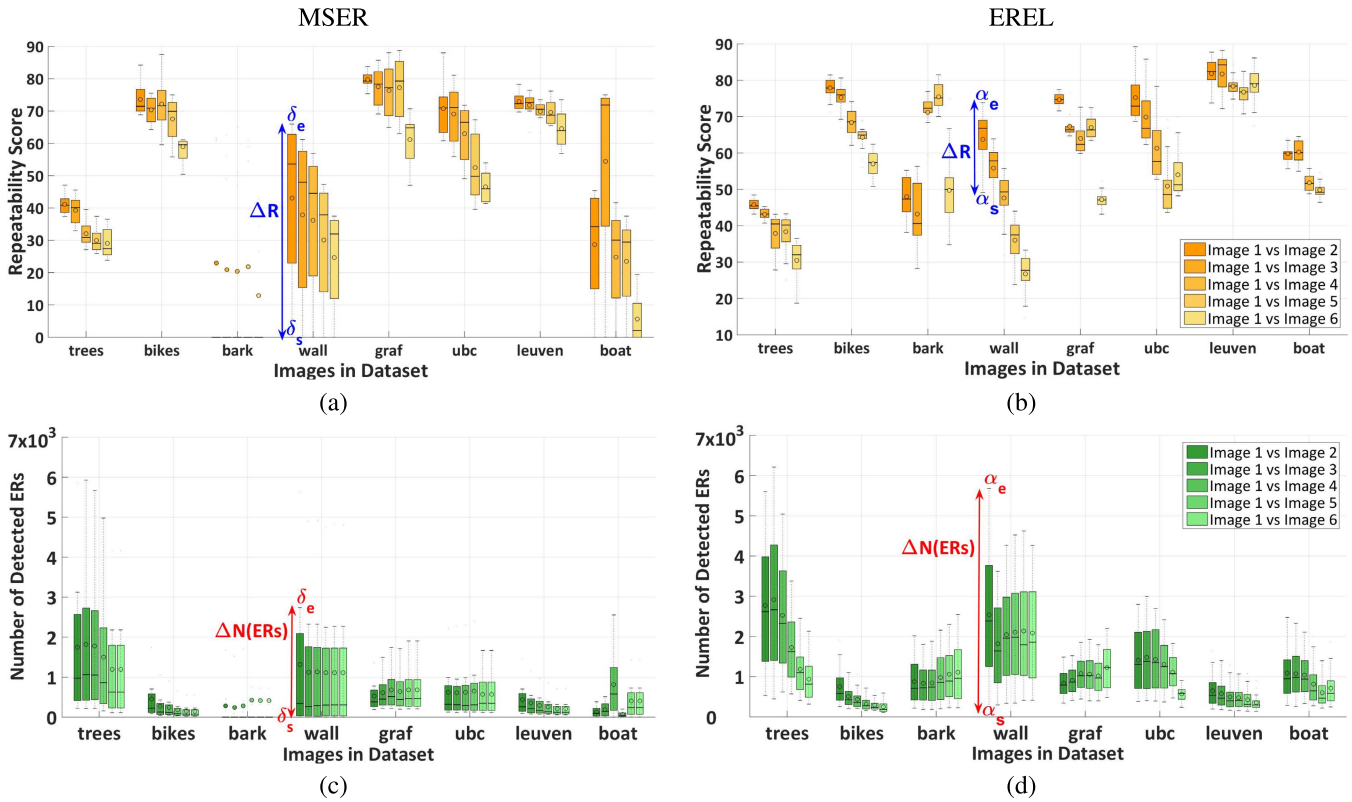

Fig. 13. Comparison of the variance of the repeatability and number of detected ERs based on various values for α (belongs to EREL) and δ (belongs to MSER) for images of [54] dataset. δ is a parameter of MSER ($\delta \in [\delta_s\ \delta_e]$, $\delta_s = 4$, $\delta_e = 32$). α is a parameter of EREL ($\alpha \in [\alpha_s\ \alpha_e]$, $\alpha_s = 0.4$, $\alpha_e = 2$). ΔR shows how much the repeatability score changes. $\Delta N(ERs)$ represents variation in the number of detected regions based on different values of δ (MSER) and α (EREL).

least N_p times, after finishing the enumeration phase. It should be noted that $N_p = \Sigma_{i=1}^{N_{ERs}} Area(ER_i)$, is the total number of extracted pixels by the detector and N_{ERs} represents the number of obtained extremal regions. The reason why we believe that by N_p we draw a fairer comparison rather than using N, is because of the content of the input image. An image may consist of a few extremal regions (less informative) but has a high resolution, it means large N. A method may result in lower number of regions, so it can process a lower number of pixels and achieves a lower elapsed time. Accordingly, if we take the total number of extracted pixels into consideration, an unbiased comparison can be drawn. Therefore, we plotted the second graph in Fig. 12(b). As it is illustrated in Fig. 12(b), the average EREL process time for each extracted pixel is lower than the Nister version of MSER [39]. Not to mention that, EREL reports all coordinates of the pixels belonging the regions in addition to the parameters of the fitted ellipse.

V. Discussion

We conducted four tests on all image sets of Oxford dataset [10] to show the performance of our proposed method, EREL. We first tested the repeatability of the detected regions by EREL against MSER and then to study how the method reacts in the real-world applications, we extracted features from the detected regions by SURF [55] and measured its recall rate. Both of the tests confirmed that EREL detection can improves the performance of an application. We also studied the presence of noise, by contaminating the images in the dataset and evaluating the repeatability of the extracted regions and measure the recall value of the descriptor. It showed that MSER detection is adversely affected by the presence of noise, however, as EREL has prior information about the edges, it detects extremal regions more robustly.

Now, we discuss an important parameter of the EREL, α, which has been denoted in Eq. (2) and compare its consequences on the performance and the number of detected regions with the most important parameters of MSER, δ. Parameter δ actually determines the stability range of a region. The higher the δ is, the lower number of regions with greater degree of stability are detected. So, it can be implied that it is not possible to adjust δ to the type of the application. The best tuned values for δ in the literature [60] are 5 or 10. On the other hand, the parameter of EREL, α, accepts various values well. It usually ranges between 0 and 2.5. For $\alpha = 0$, the method detects a huge number of MGMs, but most of them are useless. For $\alpha >= 2.5$, the method detects almost no MGMs. Particularly, α shows what rate of high informative pixels can participate in the process of detecting MGMs. The lower the α is, the weaker edge points participate in the detection phase which results in having regions that cover most parts of the image. In contrast, higher values of α, help the method to select MGMs from points located on stronger edges. Therefore, EREL can be adapted based on the application in a way that its performance is not affected by changing the value of α. Fig. 13 demonstrates this robustness.

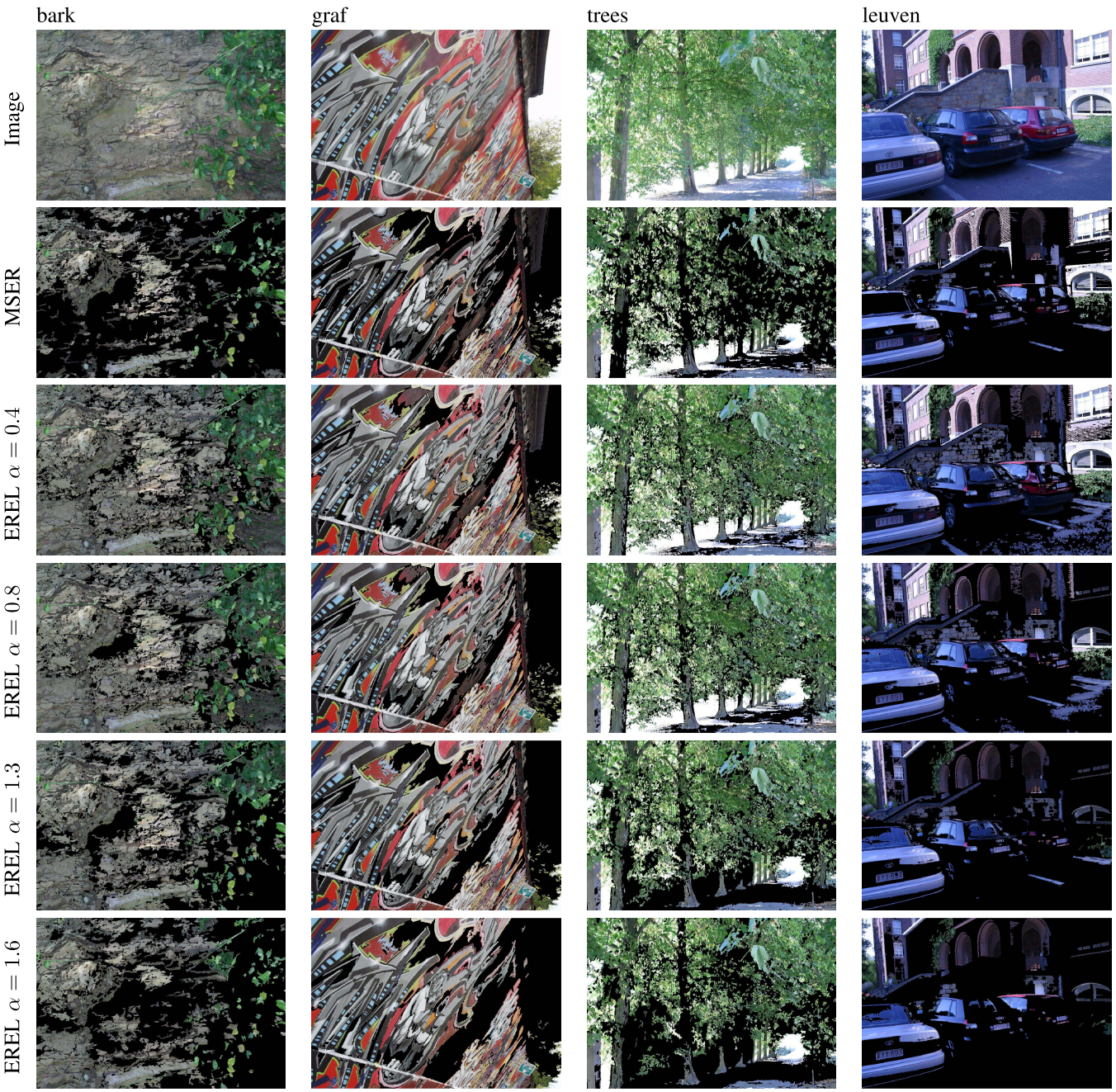

Fig. 14. Reconstruction of the images by placing pixels from the detected regions in an empty image, for images of [54] dataset based on different values for α. Note that the value for parameter δ of MSER is 10. Black pixels show that no pixel extracted from that point.

We evaluate EREL and MSER based on the various values for α ($\alpha \in [\alpha_s\ \alpha_e]$, $\alpha_s = 0.4$, $\alpha_e = 2$) and δ ($\delta \in [\delta_s\ \delta_e]$, $\delta_s = 4$, $\delta_e = 32$) for all image set in the Oxford dataset [18]. The first row of Fig. 13 ((a) and (b)) shows the change of repeatability score (ΔR) for different values of parameters of both methods. Obviously, $E[\Delta R_{MSER}] > E[\Delta R_{EREL}]$, where the function $E()$ represents the average value. What we mean is that changes in value of δ decrease the repeatability significantly, however changes in value of α does not considerably affect repeatability and keeps it in a stable condition. We conclude by a similar logic about the second row of the Fig. 13. The second row ((c) and (d)) demonstrates the number of detected regions. The maximum and minimum number of regions for MSER and EREL can be observed in Fig. 13. It can be seen that change in value of δ actually does not affect the number of ERs (Fig. 13(c)). It means that high percent of ERs are similar together, so there is no chance to produce different kinds of ERs by changing δ according to a specific application. In contrast, manipulating α can result in higher variation in the number of detected regions that not only provide a tool for adjusting the final detected regions to a specific application, but also enhances the coverage of image by detected features. To show the actual coverage of the method, we construct images by placing the pixels of the detected extremal regions into an empty image for both MSER and EREL. Fig. 14 illustrates the results for various values

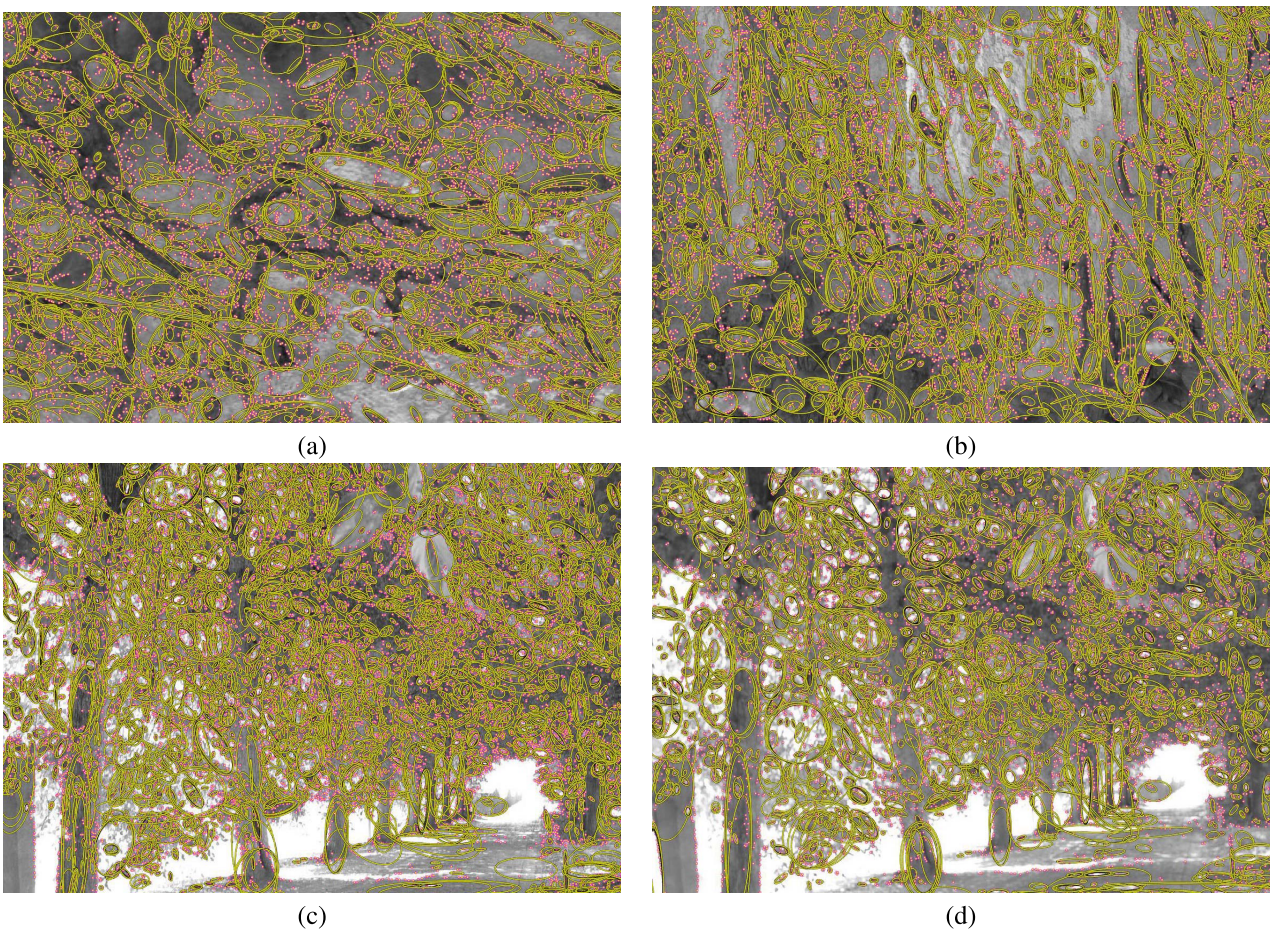

Fig. 15. The regions and MGMs found by the proposed EREL method in two pairs of images from [1] dataset. (a) "bark" image 1. (b) "bark" image 2. (c) "trees" image 1. (d) "trees" image 4.

of α. The high coverage of the EREL can be specified by comparing the second row of the Fig. 14 with other rows.

The final detected regions by EREL, are selected from the same finite set of extremal regions that MSER chooses its extremal regions from. Fig. 15 shows the output of the system for two image pairs from the Mikolajczyk dataset of [1]. There are several advantages that the general idea of EREL presents over MSER, including:

- Contribution of the edge points to the method
- Constant time calculation of the global criterion
- Having a prior knowledge to select extremal regions before the inception of the union-find algorithm
- High adaptability to the type of the application
- Linear time complexity to report both distinguished regions and measurement regions
- High repeatability rates for various transformations
- High recall rates for various transformation
- Robust performance for various transformations at the presence of noise

VI. Conclusion

This paper has introduced a novel algorithm for invariant region detection. This algorithm is inspired by the well-known Maximally Stable Extremal Regions (MSER) algorithm which detects stable repeatable regions through three steps of extremal regions enumeration, obtaining region stability criterion, and cleaning up. MSER is proven to be very useful and efficient, but, it uses no information about the boundaries of the regions. We have shown in this paper that including such information in the process of finding the extremal regions not only eliminates the need for the rather complicated step of regions enumerations and the cleaning up step of MSER, but also results in a region detector that has linear time complexity and outperforms MSER. This has been proven through experimental results on the popular benchmark dataset of Mikolajczyks in [1] which imposes different image degradations, such as blur, viewpoint change, scale change, and JPEG compression, to its image sequences. We have evaluated our proposed method based on several criterion including repeatability score for standards overlap errors (40%) and smaller ones (to show the accuracy of detected regions) in a similar condition with MSER features. Employing EREL in real-world applications, needs descriptor extraction from the detected regions. To show that our proposed method works properly in such cases we have extracted SURF [55] descriptor for EREL and MSER, and have compared their performance. The results of the comparison have shown that the performance of the EREL is either superior or close to

MSER. The robustness of our proposed method has also been studied under noisy conditions. Both the detector performance and the performance of its extracted descriptors on noisy images, have been evaluated. The obtained results show that our proposed method is more robust to noise than MSER. The coverage of EREL features has been illustrated by reconstructing the images based on the pixels that contribute to extremal regions. Finally, the proposed EREL method has been compared against recent versions of MSER. Overall, EREL detects accurate repeatable invariant regions and can be more efficiently employed in various applications, compared to MSER.

We are planning to extend the proposed algorithm to video sequences and utilize temporal information in our future work.

REFERENCES

[1] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.

[2] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. 9th ECCV*, 2006, pp. 490–503.

[3] L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, 2013.

[4] A. Agarwal and B. Triggs, "Hyperfeatures—Multilevel local coding for visual recognition," in *Proc. 9th ECCV*, 2006, pp. 30–43.

[5] T. Lindeberg, "Scale selection for differential operators," in *Scale-Space Theory in Computer Vision*. New York, NY, USA: Springer-Verlag, 1994.

[6] T. Lindeberg, *Scale-Space Theory in Computer Vision*. New York, NY, USA: Springer-Verlag, 1993.

[7] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, 1998.

[8] B. Jähne *et al.*, Eds., *Handbook on Computer Vision and Applications*, vol. 2. Boston, MA, USA: Academic, 1999, pp. 239–274.

[9] J. Gårding and T. Lindeberg, "Direct computation of shape cues using scale-adapted spatial derivative operators," *Int. J. Comput. Vis.*, vol. 17, no. 2, pp. 163–191, 1996.

[10] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[11] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2002.

[12] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or 'how do i organize my holiday snaps?"' in *Proc. 7th ECCV*, 2002, pp. 414–431.

[13] T. Tuytelaars and L. Van Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.

[14] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in *Proc. BMVC*, 2000, pp. 1–14.

[15] T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention," *Int. J. Comput. Vis.*, vol. 11, no. 3, pp. 283–318, 1993.

[16] T. Lindeberg and J. Gårding, "Shape-adapted smoothing in estimation of 3D depth cues from affine distortions of local 2D brightness structure," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 1994, pp. 389–400.

[17] T. Lindeberg, "Image matching using generalized scale-space interest points," in *Scale Space and Variational Methods in Computer Vision*. Berlin, Germany: Springer-Verlag, 2013.

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[19] J. Matas, T. Obdrzalek, and O. Chum, "Local affine frames for wide-baseline stereo," in *Proc. 16th ICPR*, vol. 4. 2002, pp. 363–366.

[20] Y. Ning, R. Chen, and P. Xu, "Wide baseline image mosaicing by integrating MSER and hessian-affine," in *Proc. 4th Int. CISP*, 2011, pp. 2034–2037.

[21] J. Xin, X. Ma, Y. Deng, D. Liu, and H. Liu, "A new method of stereo localization using dual-PTZ-cameras," in *Intelligent Robotics and Applications*. Berlin, Germany: Springer-Verlag, 2012, pp. 460–472.

[22] Š. Obdržálek and J. Matas, "Object recognition using local affine frames on distinguished regions," in *Proc. BMVC*, 2002, pp. 113–122.

[23] Š. Obdržálek and J. Matas, "Object recognition using local affine frames on maximally stable extremal regions," in *Toward Category-Level Object Recognition*. Berlin, Germany: Springer-Verlag, 2006, pp. 83–104.

[24] M. Teutsch, T. Mueller, M. Huber, and J. Beyerer, "Low resolution person detection with a moving thermal infrared camera by hot spot classification," in *Proc. CVPRW*, Jun. 2014, pp. 209–216.

[25] M. Donoser and H. Bischof, "3D segmentation by maximally stable volumes (MSVs)," in *Proc. 18th ICPR*, 2006, pp. 63–66.

[26] A. Mammeri, A. Boukerche, and G. Lu, "Lane detection and tracking system based on the MSER algorithm, Hough transform and Kalman filter," in *Proc. ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst.*, 2014, pp. 259–266.

[27] H. Deusch, J. Wiest, S. Reuter, D. Nuss, M. Fritzsche, and K. Dietmayer, "Multi-sensor self-localization based on maximally stable extremal regions," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 555–560.

[28] X. Sun, C. M. Christoudias, V. Lepetit, and P. Fua, "Real-time landing place assessment in man-made environments," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 211–227, 2014.

[29] P. Ewerling, A. Kulik, and B. Froehlich, "Finger and hand detection for multi-touch interfaces based on maximally stable extremal regions," in *Proc. ACM Int. Conf. Interact. Tabletops Surf.*, 2012, pp. 173–182.

[30] J. Guo, H. Sun, C. Zhu, and S. Xiao, "Multispectral remote sensing image registration based on maximally stable extremal regions," *Proc. SPIE*, vol. 7494, pp. 749412-1–749412-6, Oct. 2009, doi: 10.1117/12.832949.

[31] L. Liu, H. Y. Tuo, T. Xu, and Z. L. Jing, "Multi-spectral image registration and evaluation based on edge-enhanced MSER," *Imag. Sci. J.*, vol. 62, no. 4, pp. 228–235, 2014.

[32] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate Web image search," in *Proc. CVPR*, Jun. 2009, pp. 25–32.

[33] M. Okade and P. K. Biswas, "Improving video stabilization using multi-resolution MSER features," *IETE J. Res.*, vol. 60, no. 5, pp. 373–380, 2014.

[34] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. 18th IEEE ICIP*, Sep. 2011, pp. 2609–2612.

[35] M. Opitz, M. Diem, S. Fiel, F. Kleber, and R. Sablatnig, "End-to-end text recognition using local ternary patterns, MSER and deep convolutional nets," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst. (DAS)*, 2014, pp. 186–190.

[36] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Proc. 13th ECCV*, 2014, pp. 497–511.

[37] E. Murphy-Chutorian and M. M. Trivedi, "N-tree disjoint-set forests for maximally stable extremal regions," in *Proc. BMVC*, 2006, pp. 739–748.

[38] F. Kristensen and W. J. MacLean, "Real-time extraction of maximally stable extremal regions on an FPGA," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 165–168.

[39] D. Nistér and H. Stewénius, "Linear time maximally stable extremal regions," in *Proc. 10th ECCV*, 2008, pp. 183–196.

[40] F. Fraundorfer, M. Winter, and H. Bischof, "MSCC: Maximally stable corner clusters," in *Image Analysis* (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2005, pp. 45–54.

[41] P.-E. Forssen, "Maximally stable colour regions for recognition and matching," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[42] M. Perdoch, J. Matas, and S. Obdrzalek, "Stable affine frames on isophotes," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[43] L. Cheng, J. Gong, X. Yang, C. Fan, and P. Han, "Robust affine invariant feature extraction for image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 246–250, Apr. 2008.

[44] W. Zhang, Q. M. J. Wu, G. Wang, X. You, and Y. Wang, "Image matching using enclosed region detector," *J. Vis. Commun. Image Represent.*, vol. 21, no. 4, pp. 271–282, 2010.

[45] L. Ronghua and M. Huaqing, "Multi-scale maximally stable extremal regions for object recognition," in *Proc. IEEE Int. Conf. Inf. Autom.*, Jun. 2010, pp. 1799–1803.

[46] M.-M. Zhang, Z.-M. Li, H.-H. Bai, and Y. Sun, "Robust image salient regional extraction and matching based on DOGSS-MSERs," *Optik, Int. J. Light Electron Opt.*, vol. 125, no. 3, pp. 1469–1473, 2014.

[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[48] P.-E. Forssen and D. G. Lowe, "Shape descriptors for maximally stable extremal regions," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[49] J. Milnor, *Morse Theory* (Annals of Mathematic Studies AM-51). Princeton, NJ, USA: Princeton Univ. Press, 1963.

[50] R. Kimmel, C. Zhang, A. M. Bronstein, and M. M. Bronstein, "Are MSER features really interesting?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2316–2320, Nov. 2011.
[51] M. Faraji, J. Shanbehzadeh, K. Nasrollahi, and T. B. Moeslund, "Extremal regions of extremum levels," in *Proc. IEEE Signal Process. Soc. Int. Conf. Image Process. (ICIP)*, Sep. 2015.
[52] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Trans. Syst., Man, Cybern.*, vol. 8, no. 8, pp. 630–632, 1978.
[53] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. Comput. Vis.*, vol. 37, no. 2, pp. 151–172, 2000.
[54] K. Mikolajczyk and C. Schmid, "Comparison of affine-invariant local detectors and descriptors," in *Proc. 12th Eur. Signal Process. Conf*, 2005, pp. 1729–1732.
[55] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2006, pp. 404–417.
[56] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[57] Y. Xu, P. Monasse, T. Geraud, and L. Najman, "Tree-based morse regions: A topological approach to local feature detection," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5612–5625, Dec. 2014.
[58] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Dec. 2001, pp. I-511–I-518.
[59] R. Sedgewick and K. Wayne, *Algorithms*, 4th ed. Reading, MA, USA: Addison-Wesley, 2011.
[60] A. Vedaldi and B. Fulkerson. (2008). *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. [Online]. Available: http://www.vlfeat.org/

Mehdi Faraji receives the B.Sc. degree in computer engineering from the Islamic Azad University of Qazvin, and the M.Sc. degree in artificial intelligence from Kharazmi University, in 2015. His research interest is mainly feature detection and description, which are early steps of a many computer vision applications. He is also interested in cognitive science and interdisciplinary fields, such as visual neuroscience and visual perception.

Jamshid Shanbehzadeh received the B.Sc. and M.Sc. degrees from Tehran University, Iran, and the Ph.D. degree from Wollongong University, Australia. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Kharazmi University, Iran. He has authored over 200 papers on image processing and machine vision. His current activity is on image retrieval, scene understanding, image description, and image compression. He received the Best Teaching and Research Award from Kharazmi university.

Kamal Nasrollahi received the M.Sc. degree in computer engineering and electrical engineering from the Amirkabir University of Technology, in 2007, and the Ph.D. degree in computer engineering and electrical engineering from Aalborg University, Denmark, in 2010, with a focus on computer vision. He is currently an Associate Professor with the Visual Analysis of People Laboratory, Aalborg University. He has been involved in five (inter)national research projects. His research interests include facial analysis systems, biometrics recognition, soft biometrics, and inverse problems. He has won an IEEE Conference Best Paper Award.

Thomas Baltzer Moeslund is currently the Head of the Visual Analysis of People Laboratory with Aalborg University, and the Media Technology Section, Aalborg, Denmark. His research is focused on all aspects of automatic analysis of images and video data. He has been involved in 19 (inter)national research projects. He performs editorial duties for four international journals. He has been the Co-Chair of 17 international conferences/workshops/tutorials. His awards include a Most Cited Paper Award in 2009, a Teacher of the Year Award in 2010, Northern Jutland University–Foundation Innovation Award in 2013, and best paper awards in 2010, 2012, and 2014.