Pattern Recognition Letters 74 (2016) 9–16

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

On the completeness of feature-driven maximally stable extremal regions☆

Pedro Martins[a,*], Paulo Carvalho[a], Carlo Gatta[b]

[a] CISUC, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
[b] Computer Vision Center, Autonomous University of Barcelona, 08193 Bellaterra, Spain

ARTICLE INFO

Article history:
Received 18 June 2015
Available online 22 January 2016

Keywords:
Local features
Completeness
Maximally Stable Extremal Regions

ABSTRACT

By definition, local image features provide a compact representation of the image in which most of the image information is preserved. This capability offered by local features has been overlooked, despite being relevant in many application scenarios. In this paper, we analyze and discuss the performance of feature-driven Maximally Stable Extremal Regions (MSER) in terms of the coverage of informative image parts (completeness). This type of features results from an MSER extraction on saliency maps in which features related to objects boundaries or even symmetry axes are highlighted. These maps are intended to be suitable domains for MSER detection, allowing this detector to provide a better coverage of informative image parts. Our experimental results, which were based on a large-scale evaluation, show that feature-driven MSER have relatively high completeness values and provide more complete sets than a traditional MSER detection even when sets of similar cardinality are considered.

1. Introduction

Local image feature detection has been a prolific research topic in the fields of computer vision and image analysis, mostly due to the fundamental role it plays in a number of prominent tasks. Local feature detection is often found as the first step of effective algorithms targeted at solving a diversity of problems such as wide-baseline stereo matching, camera calibration, image retrieval, and object recognition. Using a sparse set of locally salient and potentially overlapping image parts – the so-called image features – offers two immediate advantages: (i) the existence of many and possibly redundant patches ensures robustness; (ii) by keeping only informative image parts, a compact image representation is constructed and, subsequently, the amount of data for further processing is reduced. Depending on the application domain, there are other properties that local features should exhibit. For example, for matching tasks, it is fundamental to have repeatable and accurate features. That is, the detector should accurately respond to the "the same" features on two different images of the same scene, regardless of the underlying image transformation. Additionally, features should be distinctive, i.e., the patterns in the immediate surroundings of the features should show a significant degree of variation among themselves. Such property allows local features to be easily distinguished through the use of local descriptors.

Given the importance of repeatability in a wide range of application domains, most studies on local feature detection have been focused on the design of repeatable detectors. Currently, there are various algorithms (e.g., Harris- ,Hessian-Affine [30] or HarrisZ [4]) which are able to detect features with a high repeatability rate even in the presence of severe image transformations, such as viewpoint changes.

The introduction of robust local descriptors (e.g., SIFT [25], SURF [2], or sGLOH [3]) has contributed to set a new paradigm in local feature detection. Besides matching patches on an individual basis, the combination of local descriptors and local features has enabled the construction of robust image representations [39], which is particularly useful to solve problems in which a semantical interpretation is involved, such as the tasks of recognizing objects, classifying scenes, or retrieving semantically equivalent images. Unlike repeatability, the study of robust and compact image representations by means of local features has been overlooked. This could be partially explained by the success of dense sampling-based representations [5,13,36,40] in object and scene recognition, which is a simpler strategy that uses local descriptors densely sampled on a regular grid.

While dense sampling is a well-established and successful strategy for object and scene recognition, there are other application domains, namely emerging ones, that could benefit from robust and simultaneously compact (sparse) image representations

☆ This paper has been recommended for acceptance by Punam Kumar Saha.

* Corresponding author. Tel.: +351 967983202.
E-mail address: pjmm@dei.uc.pt (P. Martins).

http://dx.doi.org/10.1016/j.patrec.2016.01.003
0167-8655/

via local features. The development of inexpensive cameras and storage has boosted the creation of significantly large image and video databases. In such databases, search and retrieval mechanisms only make sense if they are efficiently performed [1,26]. The use of local-feature based robust image representations might be a starting point to ensure such efficiency. This scenario emphasizes the importance of analyzing the completeness of features as well as the complementarity between different types of features. Here, completeness means the amount of image information preserved by the local features [10,16], whereas complementarity reflects how different two or more types of features are.

In the large-scale completeness and complementarity test performed by Dickscheid et al. [10], the Maximally Stable Extremal Regions (MSER) [29] showed remarkable overall results: they had significantly higher completeness values compared to other types of local features of similar sparseness. In fact, the MSER completeness values were comparable with the ones of Salient Regions [19], which appear in a substantially higher number.

In this paper, we extend the study on feature-driven Maximally Stable Extremal Regions (fMSER) [27,28], which are a derivation of MSER features and aimed at overcoming some limitations of a regular MSER detection. Our main goal is to analyze the completeness and complementarity of different fMSER. In other words, we are interested in assessing the potential suitability of this type of features for application domains requiring robust image representations via the use of local features. As a result, we present a large-scale evaluation test in which fMSER and MSER are studied in terms of completeness and complementarity.

The remainder of this paper is organized as follows. Section 2 introduces definitions and notations that will be followed throughout this document and presents the motivation behind the construction of feature-driven MSER. Section 3 covers the derivation of several instances of fMSER features from standard MSER. An evaluation of the completeness and complementarity of feature-driven MSER is presented in Section 4. Finally, conclusions and perspectives are given in Section 5.

2. Background and motivation

Boundary-related semi-local structures such as edges and curvilinear shapes are known for being more robust to intensity, color, and pose variations than typical interest points (e.g., corner points). Some local feature detectors explicitly or implicitly take advantage of this robustness by detecting stable regions from semi-local structures, such as the Edge-based Regions (EBR) detector [37,38], which is based on edge detection, or the Principal Curvature-Based Regions (PCBR) detector [9], which is based on line detection. These two examples use the detection of boundary-related structures to generate the final regions. The Maximally Stable Extremal Regions (MSER) detector [29] implicitly takes advantage of the robustness of boundary-related structures without detecting them. In fact, the MSER detector is in its essence an intensity-based region detector dealing with connected components and extracting extremal regions that are stable to intensity perturbations.

The use of boundary information in the construction of local features is not only advantageous in terms of robustness. The semantic meaningfulness of boundary information is equally relevant, as it allows local features to capture informative image parts, which contributes to the construction of intuitive object representations [21].

When the goal is to ensure a robust image representation in an efficient manner via the use of local features, the MSER detector appears as a suitable option. Extremal regions can be enumerated in almost linear time, which makes the MSER detector one of the most efficient solutions for local feature detection. Additionally, the results from the large-scale completeness evaluation performed by Dickscheid et al. [10] showed that MSER features provide a relatively robust image representation despite their sparseness.

Fig. 1. An example of standard MSER detection. Either the boundaries of objects or other contours are responsible for delineating MSER features. For a better visualization, the original MSER were replaced by fitting ellipses.

Feature-driven MSER [28] were initially proposed as an attempt to overcome the typical shortcomings of a standard MSER detection, namely the lack of robustness to blur, the reduced number of regions, and the biased preference towards round shapes [20]. By detecting more regions on informative parts of the image, feature-driven MSER represent an improvement over standard MSER in terms of completeness.

2.1. Maximally stable extremal regions

Affine covariant regions can be derived from extremal regions. In the image domain, an extremal region corresponds to a connected component whose corresponding pixels have either higher or lower intensity than all the pixels on its boundary. Extremal regions hold two important properties: the set of extremal regions is closed under continuous transformations of image coordinates as well as monotonic transformations of image intensities. The Maximally Stable Extremal Regions detector responds to extremal regions that are stable with respect to intensity perturbations (see Fig. 1). For a better understanding of the MSER detector, we introduce the formal definitions of connected component and extremal regions [33].

A connected component (or region) $\mathcal{Q}$ in $\mathcal{D}$ is a subset of $\mathcal{D}$ for which each pair of pixels $(\mathbf{p}, \mathbf{q}) \in \mathcal{Q}^2$ is connected by a path in $\mathcal{Q}$, i.e., there is a sequence $\mathbf{p}, \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m, \mathbf{q} \in \mathcal{Q}$ such that $\mathbf{p} \sim \mathbf{a}_1, \mathbf{a}_1 \sim \mathbf{a}_2, \ldots, \mathbf{a}_m \sim \mathbf{q}$, where $\sim$ is the equivalence relation defined by $(\mathbf{p} \sim \mathbf{q}) \iff \max\{|p_1 - q_1|, |p_2 - q_2|\} \leq 1$ (8-neighborhood).

We define the boundary of a region $\mathcal{Q}$ as the set $\partial\mathcal{Q} = \{\mathbf{p} \in \mathcal{D} \setminus \mathcal{Q} : \exists \mathbf{q} \in \mathcal{Q} : \mathbf{p} \sim \mathbf{q}\}$. A connected component $\mathcal{Q}$ in $\mathcal{D}$ is an extremal region if $\forall \mathbf{p} \in \mathcal{Q}, \mathbf{q} \in \partial\mathcal{Q} : I(\mathbf{p}) < I(\mathbf{q})$ or $I(\mathbf{p}) > I(\mathbf{q})$.

Let $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \mathcal{Q}_{i+1} \ldots$ be a sequence of extremal regions such that $\mathcal{Q}_k \subset \mathcal{Q}_{k+1}$, $k = 1, 2, \ldots$. We say that $\mathcal{Q}_i$ is a maximally stable extremal region if and only if the stability criterion

$$\rho(k, \Delta) = \frac{|\mathcal{Q}_{k+\Delta} \setminus \mathcal{Q}_k|}{|\mathcal{Q}_k|} \tag{1}$$

attains a local minimum at i, where Δ is a positive integer denoting the stability threshold. As the area ratios are preserved under affine transformations, ρ is invariant with respect to affine transformations. Consequently, MSER features become covariant with this type of geometric transformations.

2.1.1. Advantages

As already mentioned in the introductory section, MSER tend to provide a good coverage of informative image parts, despite

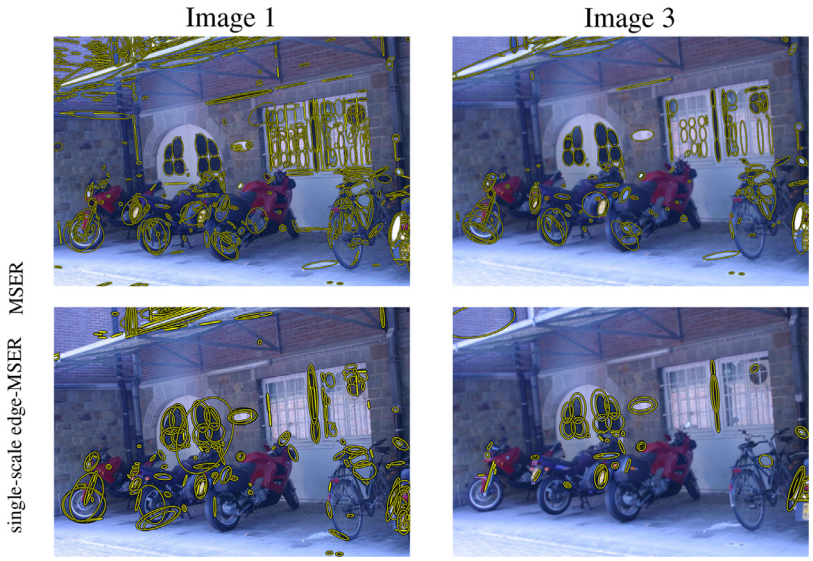

Fig. 2. Using the gradient magnitude to detect feature-driven MSER on images 1 and 3 of Bikes sequence [31]. Top row: standard MSER; bottom row: feature-driven MSER.

their typical sparseness. In addition, MSER can be efficiently detected. Extremal regions can be enumerated in almost linear time, which is a significant advantage of the MSER detector over other affine covariant regions detectors. Note that other popular algorithms for the detection of affine covariant features are usually more computationally expensive. For example, Harris-Affine and Hessian-Affine algorithms [30] detect initial points in linear time; however, this task is complemented with an automatic scale selection and an affine shape adaptation algorithm whose complexity is $\mathcal{O}(p(s+k))$, where p is the number of initial points, s is the number of scales, and k is the number of iterations required for shape adaptation [31].

2.1.2. Disadvantages

In the large-scale comparative study on affine covariant regions performed by Mikolajczyk et al. [31], MSER and Hessian-Affine features showed higher repeatability scores. However, the MSER detector showed an inconsistent performance: blurred sequences of images as well as textured sequences produced less repeatable features. The low repeatability scores in the above-mentioned conditions is a well known downside of MSER detection. The sensitiveness to image blur can be explained by the undermining effect that blur has on the stability criterion: by applying different levels of blur, we change the area of extremal regions. Additionally, as the blurring effect increases, the number of extremal regions decreases. As for textured scenes, they are not a suitable domain for MSER detection since intensity perturbations cause an irregular area variation of extremal regions in busy parts of the image. The preference for round shapes is another downside of this detector [20].

2.1.3. Derivations and applications

Over the years, different derivations of the MSER detector have been proposed. For example, the MSER algorithm has been modified to deal with volumetric [11] and color images [14]. Some authors have proposed efficiency enhancements [12,32,34] or a muti-resolution version [15].

With regard to applications, MSER features have been used in several heterogeneous tasks, such as matching [15], tracking [12], road traffic sign recognition [17], or text detection [8], among others.

3. Feature-driven maximally stable extremal regions

The ideal image for the MSER detector is the one that is well-structured, with uniform regions separated by strong intensity changes [39]. A feature-driven MSER is a region resulting from an MSER detection on a saliency map in which boundary-related features are highlighted. To perform robust text detection, [8] followed a similar strategy, which consisted of performing an edge-enhanced MSER detection based on the Canny edge detector [7]. In our case, a simple and straightforward highlighting such as a single-scale edge highlighting will not be advantageous. To illustrate this, we can think of a measure of the edge response such as the gradient magnitude computed at a single scale. Fig. 2 depicts a feature-driven MSER detection based on a single-scale gradient magnitude under the presence of blur induced by de-focus. In the example given, one can observe that the proposed feature highlighting neither reduces the typical sparseness of standard MSER nor improves the robustness to blur. An edge response computed at a single scale is rarely sufficient to capture the presence of all the boundaries in a scene. In addition, the induced blur weakens the edge response.

Another alternative is to consider feature highlighting via anisotropic diffusion [35]. With an anisotropic diffusion, there is the clear advantage of preserving details and simultaneously blurring noise, which appears to be ideal for constructing a suitable domain for MSER detection. Nonetheless, anisotropic diffusion has a relatively high computational complexity. One of our aims is that feature-driven MSER detection preserves the major advantages of a standard MSER detection, which means that the computational complexity of the whole method should be kept low. Such requirement hinders the use of non-linear scale spaces, even if we consider more efficient solutions to anisotropic diffusion [e.g., 18].

Given the constraints in terms of computational complexity, we opt for a linear (Gaussian) scale-space representation in which we highlight boundary-related features and simultaneously delineate smooth transitions at the boundaries. On one hand, the feature highlighting allows us to have well-defined boundaries, which tends to increase the number of stable extremal regions with respect to intensity perturbations. On the other hand, the existing smoothness at the boundaries attenuates the undermining effect that blurring has over the MSER detector.

The saliency maps in which the boundary-related features are highlighted are given by the following equation:

$$F(\mathbf{x}) = \sum_{i=1}^{N} \sigma_i^k S(\mathbf{x}, \sigma_i), \tag{2}$$

where N denotes the number of scales, σ_i is the ith scale, with $\sigma_i = \xi\sigma_0^{i-1}$ ($\xi \in \mathbb{R}^+$, $\sigma_0 > 1$), i.e, scale varies in a geometrical sequence, $S(\mathbf{x}, \sigma_i)$ measures the response of a boundary-related feature at pixel $\mathbf{x}$ and scale σ_i and k is the parameter required to construct normalized scale-space derivatives [24].

The feature-driven MSER detector can be seen as a hybrid between the MSER algorithm and the PCBR detector, since we make use of structural information (shapes) to define suitable domains for MSER detection. The idea is to combine the advantages of PCBR detection (explicit use of structural information) with the advantages of MSER detection (computational efficiency, repeatability, and accuracy) and simultaneously overcome some of the major limitations of the latter, namely the lack of robustness to blurring and the biased preference for round shapes. We present three different ways of highlighting boundary-related features, namely edges and lines.

3.1. Edge highlighting

To highlight edges, we explore two differential-based measures. For our purpose, there is not a clear advantage of one measure over the other. In fact, even though they highlight the same structure, they can be combined in the construction of sets of complementary fMSER.

3.1.1. Edge highlighting (version 1)

Let $L(:, \sigma)$ be a smoothed version of an image I by means of a Gaussian kernel G at the scale σ, i.e., $L(\mathbf{x}, \sigma) = G(\sigma) * I(\mathbf{x})$. The edge strength can be found by measuring the gradient magnitude,

$$S_g(\mathbf{x}, \sigma) = |\nabla L(\mathbf{x}, \sigma)| \stackrel{\text{def}}{=} \sqrt{L_x^2(\mathbf{x}, \sigma) + L_y^2(\mathbf{x}, \sigma)}, \tag{3}$$

where L_x and L_y denote the first-order partial derivatives of L in the x and y directions, respectively.

3.1.2. Edge highlighting (version 2)

The second measure is based on the eigenvalues of the structure tensor matrix. Near edges, at least one of the eigenvalues is large, while in flat areas, both values tend to zero. This suggests the use of the maximum eigenvalue – λ_2 – as a measure of edge strength. However, the range of values found for the maximum eigenvalue is considerably wide. To obtain a more balanced output, the natural logarithm is used [22]:

$$S_\mu(\mathbf{x}, \sigma) = \max\{0, \log(\lambda_2(\mu(\mathbf{x}, \sigma)))\}, \tag{4}$$

where $\mu(\mathbf{x}, \sigma)$ is the structure tensor matrix computed at pixel $\mathbf{x}$ and scale σ:

$$\mu(\mathbf{x}, \sigma) \stackrel{\text{def}}{=} G(\sigma) * \begin{bmatrix} L_x^2(\mathbf{x}, s\sigma) & L_xL_y(\mathbf{x}, s\sigma) \\ L_xL_y(\mathbf{x}, s\sigma) & L_y^2(\mathbf{x}, s\sigma) \end{bmatrix}. \tag{5}$$

In (5), the parameter s is a real value in the range]0, 1] required to keep the derivation scale lower than the integration scale.

3.2. Line highlighting

To highlight curvilinear structures, the principal curvature [9] is used. The measure for line highlighting is derived from the Hessian matrix:

$$\mathcal{H}(\mathbf{x}, \sigma) \stackrel{\text{def}}{=} \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{xy}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix}, \tag{6}$$

where L_{xx}, L_{xy} and L_{yy} are the second order partial derivatives of L, a Gaussian smoothed version of image I. The principal curvature, which highlights curvilinear structures is either given by

$$P_{max}(\mathbf{x}, \sigma) = \max(0, \lambda_2(\mathcal{H}(\mathbf{x}, \sigma))) \tag{7}$$

or

$$P_{min}(\mathbf{x}, \sigma) = \min(0, \lambda_1(\mathcal{H}(\mathbf{x}, \sigma))), \tag{8}$$

where λ_1 and λ_2 denote the minimum and maximum eigenvalues, respectively. Note that (7) and (8) respond to complementary structures: the former responds to dark lines on a brighter background, whereas the latter detects brighter lines on a dark background. Our line highlighting measure uses the principal curvature measure to detect darker lines on a bright background. However, the measures defined in (7) and (8) can be used interchangeably:

$$S_{\mathcal{H}}(\mathbf{x}, \sigma) = \max\{\lambda_2(\mathcal{H}(\mathbf{x}, \sigma)), 0\}. \tag{9}$$

3.3. Saliency maps

In Fig. 3, we depict the proposed feature highlighting using several scales. These three instances of the proposed feature highlighting provide well-defined boundaries accompanied with smooth transitions. It is readily seen that any of the three saliency maps preserves the structural information of the image and adds some smoothness to the scene (the number of scales and the number of scales play an important role in the definition of blur). While the two types of edge highlighting mainly capture and accentuate objects boundaries, the proposed line highlighting provides a clearer structural sketch of the scene [9]. At first glance, the blur induced by the feature highlight is an undesirable property, as it tends to reduce the localization accuracy of the objects. However, this type of blur becomes advantageous in recognition tasks, namely object class recognition in which intra-class variations are desirable.

4. Comparative evaluation

As already mentioned, we are primarily interested in analyzing the completeness and complementarity of fMSER features. In our previous works, we showed that these regions appear in higher number than standard MSER and since they tend to cover salient image elements, fMSER features will be more complete. However, our previous studies have never been focused on analyzing the completeness of fMSER and MSER features when there is not a discrepancy in the number of detected regions. Here, we will perform a large-scale completeness and complementarity evaluation using a similar number of regions.

We followed the evaluation protocol proposed by Dickscheid et al. [10] to measure the completeness as well as the complementary of feature-driven MSER. This comparative study is complemented with the analysis of the coverage of globally salient elements.

The dataset used in the evaluation (Fig. 4) comprised four categories of natural scenes [13,23], the Brodatz texture collection [6], and a set of aerial images. This set of image categories corresponds to the dataset used by Dickscheid et al. [10], with the exception of a collection of cartoon images, which was not made publicly available.

To measure completeness, Dickscheid et al. [10] compute an entropy density $p_H(\mathbf{x})$ based on local image statistics and a feature coding density $p_c(\mathbf{x})$ derived from a given set of features (see Appendix A). The (in)completeness measure corresponds to the Hellinger distance between the two densities:

$$d_H(p_H, p_c) = \sqrt{\frac{1}{2}\sum_{\mathbf{x}\in\mathcal{D}}(\sqrt{p_H(\mathbf{x})} - \sqrt{p_c(\mathbf{x})})^2}, \tag{10}$$

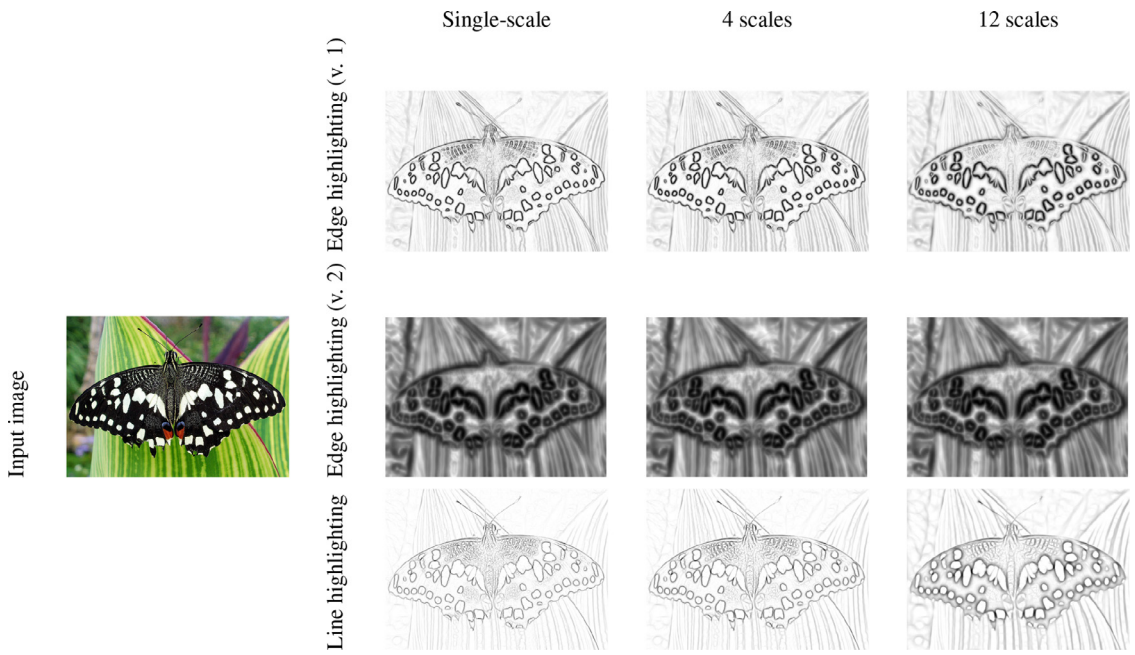

Fig. 3. Proposed feature highlighting. Darker structures in the saliency maps are the most salient ones. The parameters ξ and σ_0 were set to 1 and $\sqrt[4]{2}$, respectively. For edge highlighting based on the eigenvalues of the structure tensor matrix, the factor s (Eq. (5)) was set to 0.5.

Fig. 4. Example images from the categories in the dataset for completeness and complementarity evaluation.

where $\mathcal{D}$ is the image domain. When p_H and p_c are very close, the distance d_H will be small, which means the set of features with a coding density p_c effectively covers the image content (the set of features has a high completeness). Such metric penalizes the use of large scales (a straightforward solution to achieve a full coverage) as well as the presence of features in pure homogeneous regions. On the other hand, it will reward the "fine capturing" of local structures or superimposed features appearing at different scales.

We built the saliency maps with $\xi = 1$, $\sigma_0 = \sqrt[4]{2}$, and $N = 12$. For the edge highlighting based on the eigenvalues of the structure tensor matrix, the size of the local (derivation) scale was set to half of the size of the integration scale ($s = 0.5$ in Eq. (5)). The stability threshold Δ was set to 7 for all the instances of the fMSER, whereas for the MSER detector, this parameter was set to 5. The minimum and maximum region areas were set to 30 and 1% of the image area, respectively. We only considered MSER and fMSER features whose ρ was lower than 1.0. These parameter settings allowed us to have a similar number of regions among the different type of features. Note that the saliency maps produced by the proposed feature highlighting allow us to capture more extremal regions than the luminance channel of an image, as shown in Fig. 5.

The implementation of the MSER detector corresponds to the code provided by [41]. For fMSER features, this code was modified to deal with images whose intensity values vary in a range different from {0,. . .,255}, since the saliency maps intensity values might be greater than 255. To compute the coding and entropy densities, we used the code provided by Dickscheid et al. [10]. The number of scales used to compute the entropy density was 6 (default value).

Fig. 6 summarizes the results of our completeness evaluation. For reference, we have also included Salient Regions results in these plots. The main conclusion to be drawn from the plots is that any instance of fMSER detection provides us a more complete set of features than the one comprised of standard MSER, even when the number of regions is similar. In all categories, with the exception of Brodatz, we observe that MSER features are the least complete regions. Among feature-driven regions, it is not clear the advantage of one instance over the other ones. The three instances provide similar completeness values, although they are slightly higher than the ones obtained with the Salient regions detector, which provides a considerably higher number of features. edge2-MSER results are particularly worth of note: they appear in lower number in the Mountain category; however, they are the most complete instance of fMSER. The results for Brodatz category are explained by the fact that it only contains highly textured images, which enables the creation of very small and duplicated extremal regions in the saliency maps. In either cases, the regions were discarded.

To assess complementarity, we constructed all possible combinations of pairs of MSER and fMSER features for the first 20 images of each category. Fig. 7 depicts the completeness values for the different combined sets of features. Overall, these combinations

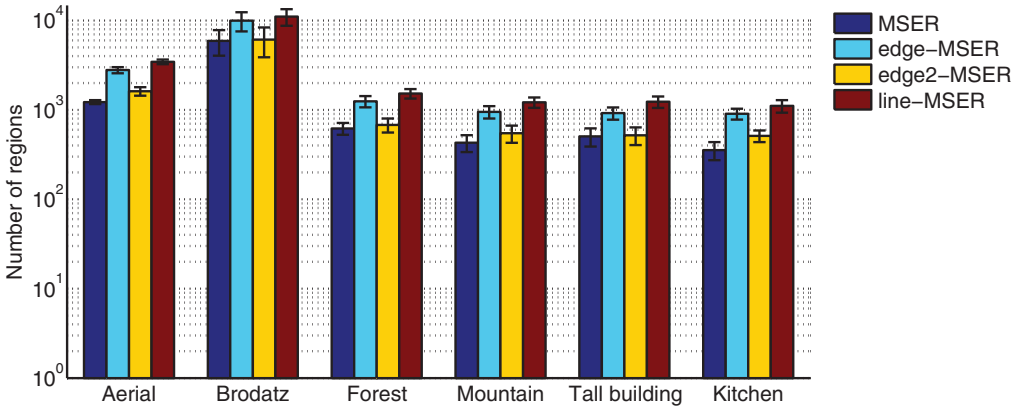

Fig. 5. Number of extremal regions provided by the saliency maps (edge-, edge2-, line-MSER) and the luminance channel (MSER) for the different categories of the dataset. Error bars indicate the standard deviation.

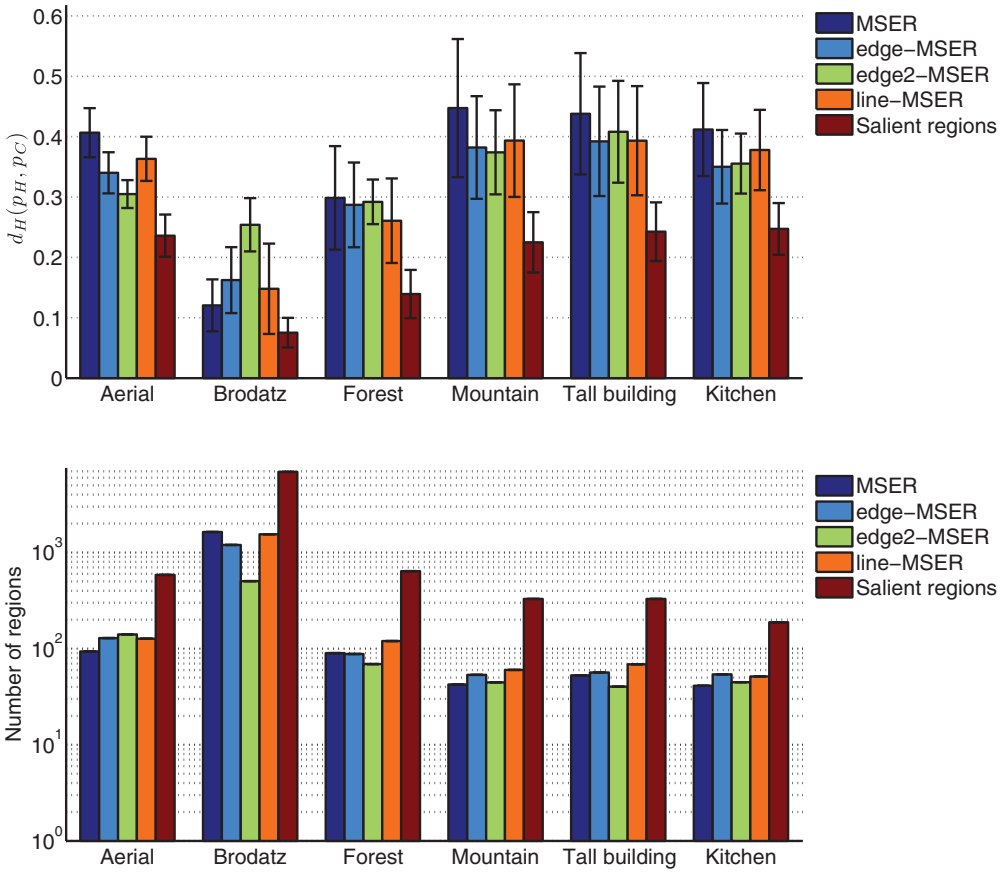

Fig. 6. Completeness results. Top row: Average dissimilarity measure $d_H(p_H, p_c)$ for the different sets of features extracted over the categories of the dataset. Bottom row: Average number of extracted features per image category. Error bars indicate the standard deviation.

are advantageous. An important observation to be drawn from the plot is that there is some redundancy between MSER and line-MSER features. Despite the latter showing a good coverage of informative parts, as seen in Fig. 6, the combination of both type of features is the least advantageous. This result is partially explained by the fact that line-MSER features are detected on a map that provides a clearer structural sketch of the image; therefore, the extracted regions will not differ too much from the ones extracted from the luminance channel. On the other hand, the combination edge2-MSER+line-MSER tends to yield the most complete results.

We complemented our evaluation with an analysis of the coverage of globally salient image parts in the first 20 images of each category. By computing the Hellinger distance between the feature coding density and a density derived from the map given by the Boolean Map-based Saliency (BMS) model [42], we were able to measure the coverage of globally salient parts. Fig. 8 summarizes these results.

For this type of coverage, fMSER features outperform MSER ones, being the exception the case of highly textured images (Brodatz). Among fMSER features, edge-MSER and line-MSER usually have the best performance. Conversely, the results show that the coverage of globally salient elements provided by edge2-MSER is slightly worse. Such fact is explained by the existence of noisier and less accurate edge maps, which can lead to extremal regions anchored at homogeneous regions.

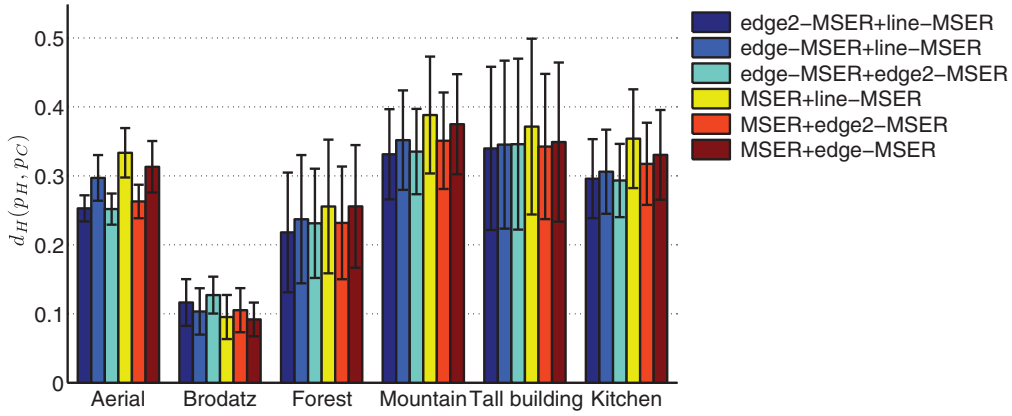

Fig. 7. Complementarity results. Average dissimilarity measure $d_H(p_H, p_c)$ for the different sets of combined features extracted over the categories of the dataset.

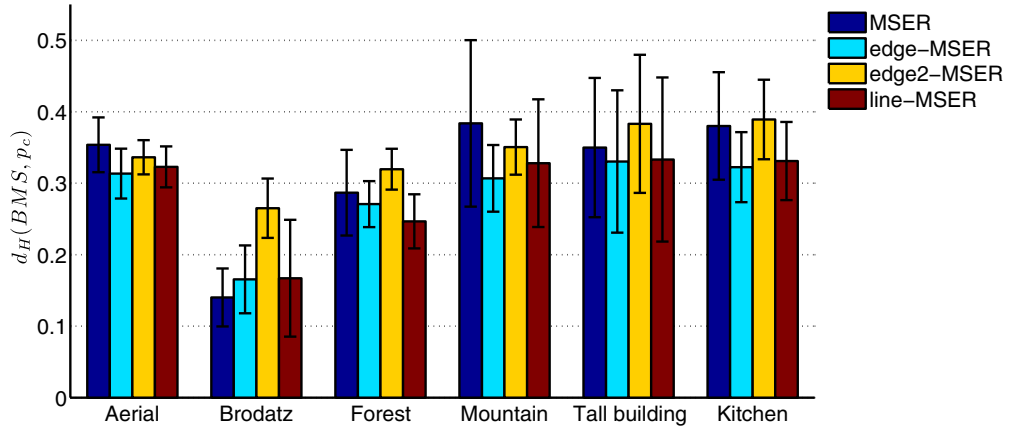

Fig. 8. Average dissimilarity measure $d_H(BMS, p_c)$ for the different sets of features extracted over the categories of the dataset.

5. Conclusions and perspectives

Feature-driven Maximally Stable Regions were initially proposed as an attempt to overcome some of the typical shortcomings of a standard MSER detection, namely the lack of robustness to blur and the reduced number of regions. This is achieved by using a particular type of saliency maps as the input image for MSER detection. In these maps, boundary-related features are simultaneously highlighted and delineated under smooth transitions. As for the number of regions, the novel features are in a substantially higher number than standard MSER ones. The study presented in this paper showed that the feature-driven regions do not only provide a better coverage of the informative content when the number of feature-driven regions is higher but also when a similar number of regions is used.

Since these regions can be obtained from different boundary-related features, it is fundamental to measure the level of complementarity among the various instances. Although they tend to provide similar regions, there is still some complementarity that can be exploited, as shown by our study.

As for the applicability of feature-driven MSER, we believe that image compression is a particularly interesting domain. This is not only due to the fact that they capture the most informative parts (robust representation) but also to the fact that an efficient algorithm is responsible for their detection. Cloud-based compression and image set compression are two current and relevant problems related to image compression where the use of fMSER features could be exploited. In both scenarios, compact robust image representations are usually obtained via the use of local descriptors. By computing such descriptors over fMSER features, one can efficiently obtain robust representations with a reduced number of regions per image.

Appendix A. Feature coding and entropy densities

The entropy density p_H is computed from local image patches with different sizes (scales). It is assumed that these patches represent a larger image, i.e., an $N \times N$ patch is part of a periodic image with period N in both directions. In addition, an image is considered to be a noisy version of a Gaussian process. From these assumptions, the entropy of an image patch g can be derived as follows:

$$H(g) = \frac{1}{2}\log_2\left(2\pi \exp\left(\frac{\det \Sigma_{gg}}{\sigma_n^2}\right)\right), \tag{A.1}$$

where Σ_{gg} represents the covariance matrix of the intensity values in g and σ_n^2 is the noise variance. The determinant of Σ_{gg} is derived from the power spectrum $P(\mathbf{u}) = |DCT(g(\mathbf{x}))|^2$, i.e., $\det \Sigma_{gg} = \prod_{\mathbf{u}\setminus\{\mathbf{0}\}} P(\mathbf{u})$, where $\mathbf{0}$ is the DC coefficient. By assuming that the powerspectrum is additively composed of the powerspectra of the signal and the noise, the following estimate of the power spectrum can be used:

$$\hat{P}(\mathbf{u}) = \max(P(\mathbf{u}) - \sigma_n^2, 0), \tag{A.2}$$

and (A.1) becomes

$$H(\mathbf{x}) = \frac{1}{2N^2}\sum_{\mathbf{u}\setminus\{\mathbf{0}\}} \max\left(\log_2\left(2\pi \exp\left(\frac{\hat{P}(\mathbf{u})}{\sigma_n^2}\right)\right), 0\right). \tag{A.3}$$

The entropy at a pixel $\mathbf{x}$ will be obtained from the patches entropy. If $H(\mathbf{x}, N)$ is the entropy of a pixel $\mathbf{x}$ based on a patch of size N, the

entropy at pixel $\mathbf{x}$ is

$$H(\mathbf{x}) = \sum_{s=1}^{S} H(\mathbf{x}, 1 + 2^s), \tag{A.4}$$

where $s \in \{1, \ldots, S\}$ denotes the scale. Finally, the density p_H is computed through normalization:

$$p_H(\mathbf{x}) = \frac{H(\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{D}} H(\mathbf{y})}. \tag{A.5}$$

The feature coding density p_c is computed for a given set of features $\mathcal{F}$. It is assumed that a feature $f \in \mathcal{F}$ can be characterized by its location $\mathbf{m}_f$ and its scale σ_f (or Σ_f in the case of affine covariant features). A Gaussian distribution spreading over the image domain is used to represent a region covered by a local feature. It is also assumed that $c(f)$ bits are required to represent a feature f. Such assumptions lead to the coding map

$$c(\mathbf{x}) = \sum_{f\in\mathcal{F}} c(f) G(\mathbf{x}, \mathbf{m}_f, \Sigma_f), \tag{A.6}$$

where G denotes an anisotropic Gaussian kernel. The final coding density p_C is the result of a normalization:

$$p_c(\mathbf{x}) = \frac{c(\mathbf{x})}{\sum_{y\in\mathcal{D}} c(\mathbf{y})}, \tag{A.7}$$

which allows us to compare both densities.

References

[1] A. Araujo, M. Makar, V. Chandrasekhar, D. Chen, S. Tsai, H. Chen, R. Angst, B. Girod, Efficient video search using image queries, in: Proceedings of 2014 IEEE International Conference on Image Processing, 2014.

[2] H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, Comput. Vis. Underst. 100 (3) (2008) 346–359.

[3] F. Bellavia, D. Tegolo, E. Trucco, Improving sift-based descriptors stability to rotations, in: Proceedings of 20th International Conference on Pattern Recognition, 2010, pp. 3460–3463.

[4] F. Bellavia, D. Tegolo, C. Valenti, Improving harris corner selection strategy, IET Comput. Vis. 5 (2) (2011) 87–96.

[5] A. Bosch, A. Zisserman, X. Muñoz, Image classification using random forests and ferns, in: Proceedings of 11th IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[6] P. Brodatz, Textures: A Photographic Album for Artists and Designers, Dover, New York, NY, USA, 1966.

[7] J. Canny, A computational approach to edge detetction, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6) (1986) 679–698.

[8] H. Chen, S.S. Tsai, G. Schroth, D.M. Chen, R. Grzeszczuk, B. Girod, Robust text detection in natural images with edge-enhanced maximally stable extremal regions, in: Proceedings of 2011 IEEE International Conference on Image Processing, 2011, pp. 2609–2612.

[9] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, L. Shapiro, Principal curvature-based region detector for object recognition, in: Proceedings of 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[10] T. Dickscheid, F. Schindler, W. Förstner, Coding images with local features, Int. J. Comput. Vis. 94 (2) (2011) 154–174.

[11] M. Donoser, H. Bischof, 3d segmentation by maximally stable volumes (MSVs), in: Proceedings of the 18th International Conference on Pattern Recognition, vol. 1, 2006, pp. 63–66.

[12] M. Donoser, H. Bischof, Efficient maximally stable extremal region (MSER) tracking, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 553–560.

[13] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 524–531.

[14] P.-E. Forssén, Maximally stable colour regions for recognition and matching, in: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[15] P.-E. Forssén, D. Lowe, Shape descriptors for maximally stable extremal regions, in: Proceedings of the 11th IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[16] W. Förstner, T. Dickscheid, F. Schindler, On the completeness of coding with image features, in: Proceedings of the British Machine Vision Conference 2009, 2009.

[17] J. Greenhalgh, M. Mirmehdi, Real-time detection and recognition of road traffic signs, IEEE Trans. Intell. Transp. Syst. 13 (4) (2012) 1498–1506.

[18] S. Grewenig, J. Weickert, A. Bruhn, From box filtering to fast explicit diffusion, in: Proceedings of the 32nd DAGM Conference on Pattern Recognition, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 533–542.

[19] T. Kadir, A. Zisserman, M. Brady, An affine invariant salient region detector, in: Proceedings of the 8th European Conference on Computer Vision, 2004, pp. 228–241.

[20] R. Kimmel, C. Zhang, A. Bronstein, M. Bronstein, Are MSER features really interesting? IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2316–2320.

[21] I. Kokkinos, A. Yuille, Scale invariance without scale selection, in: Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[22] A. Kovacs, T. Sziranyi, Improved force field for vector field convolution method, in: Proceedings of 2011 IEEE International Conference on Image Processing, 2011, pp. 2853–2856.

[23] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.

[24] T. Lindeberg, Scale-Space Theory in Computer Vision, Kluwer Academic Publishers, 1994.

[25] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[26] H. Mansour, S. Rane, P.T. Boufounos, A. Vetro, Video querying via compact descriptors of visually salient objects, in: Proceedings of 2014 IEEE International Conference on Image Processing, 2014.

[27] P. Martins, P. de Carvalho, C. Gatta, Stable salient shapes, in: Proceedings of 2012 International Conference on Digital Image Computing Techniques and Applications, 2012.

[28] P. Martins, C. Gatta, P. de Carvalho, Feature-driven maximally stable extremal regions, in: Proceedings of International Conference on Computer Vision Theory and Applications, vol. 1, 2012, pp. 490–497.

[29] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceedings of British Machine Vision Conference 2002, 2002, pp. 384–393.

[30] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L.V. Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1/2) (2005) 43–72.

[32] E. Murphy-Chutorian, M. Trivedi, N-tree disjoint-set forests for maximally stable extremal regions, in: Proceedings of British Machine Vision Conference 2006, 2006.

[33] L. Neumann, J. Matas, Real-time scene text localization and recognition, in: Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 3538–3545.

[34] D. Nistér, H. Stewénius, Linear time maximally stable extremal regions, in: Proceedings of 10th European Conference on Computer Vision, 2008, pp. 183–196.

[35] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 629–639.

[36] E. Tola, V. Lepetit, P. Fua, A fast local descriptor for dense matching, in: Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[37] T. Tuytelaars, L.V. Gool, Matching widely separated views based on affine invariant regions, Int. J. Comput. Vis. 59 (1) (2004) 61–85.

[38] T. Tuytelaars, L.V. Gool, L. D'haene, R. Koch, Matching of affinely invariant regions for visual servoing, in: Proceedings of 1999 IEEE International Conference on Robotics and Automation, 1999, pp. 1601–1606.

[39] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comput. Graph. Vis. 3 (3) (2008) 177–280.

[40] T. Tuytelaars, C. Schmid, Vector quantizing feature space with a regular lattice, in: Proceedings of Foundations 11th IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[41] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, 2008. (http://www.vlfeat.org/). (accessed 15.06.15)

[42] J. Zhang, S. Sclaroff, Saliency detection: a boolean map approach, in: Proceedings of 14th IEEE International Conference on Computer Vison, 2013, pp. 153–160.