A DATA-DRIVEN REGION DETECTOR FOR STRUCTURED IMAGE SCENES

Elena Ranguelova

Netherlands eScience Center
Amsterdam, The Netherlands
e-mail: E.Ranguelova@esciencecenter.nl

ABSTRACT

Finding correspondences between two images of the same scene, taken from different viewpoints, with semantic features is a challenging problem. This paper proposes a Data-driven Morphology Salient Regions (DMSR) approach for detecting interest regions repeatedly. A binarization algorithm creates a compact image representation that is then analyzed for saliency using morphology. DMSR has comparable performance to the renowned Maximally Stable Extremal Regions (MSER) detector on structured scenes and better invariance to lighting, blur and on a high-resolution benchmark. This is achieved via significantly fewer detected regions, leading to better scalability. DMSR is shown to be a better choice than MSER for analysis of scientific imagery in the big data era, e.g., it detects precisely meaningful regions in images used for wild-life biometrics. The paper also introduces OxFrei, a dataset for transformation-independent detection evaluation.

Index Terms— region detection, data-driven, structured scenes, morphology, scientific visual analytics

1. INTRODUCTION

The first fundamental step in numerous computer vision applications (wide baseline stereo matching, image retrieval, visual mining, etc.) is to reliably and repeatedly find the correspondence between a pair of different images of the same scene [1, 2, 3]. The class of *region detectors* find distinct (salient) regions, which correspond to the same image patches, detected independently for each viewpoint. The detectors must be *covariant* (often called *invariant*) to, usually, *affine* transformations and various photometric distortions.

While most research has focused on generic applications, the emerging fields of *animal and plant biometrics* are attracting more attention of the community [4, 5]. Computer vision is becoming a vital technology enabling the wild-life preservation efforts of ecologists in the big data era. Along with the individual or species photo-ID, the scientists wish to obtain reliable measurements of meaningful structures from images. The generic region detectors do not satisfy this need: Fig. 1, top row shows over-abundant regions often without semantics as well as missed structures, also seen in Fig. 2, (a), (b).

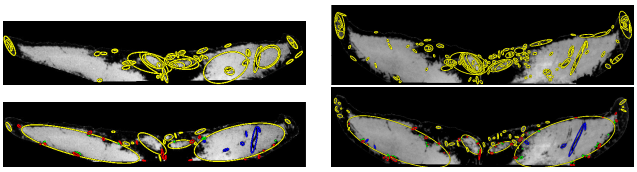
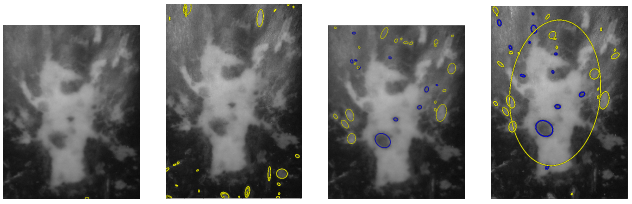


Fig. 1. Region detection on two images of the tail of the same humpback whale. Top row: MSER, bottom row: DMSR(All).



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

Fig. 2. Region detection on two images of the pineal spot of the same leatherback turtle. (a),(b): MSER, (c),(d): DMSR.

A decade ago, a performance evaluation paper by the Visual Geometry Group in Oxford compared existing region detectors [6]. A clear conclusion of the comparison was that *Maximally Stable Extremal Regions (MSER)* is the best performing detector for *structured* scenes [1]. MSER has become the de-facto standard in the field, e.g., it is in the MATLAB CVS Toolbox and OpenCV. Despite its success, the detector has several drawbacks: sensitivity to blur, producing nested and redundant regions and degrading performance with increase of image resolution [7]. Analysis in geometric scale-space showed that the formulation of the stability criterion makes MSER prefer regular shapes [8].

Many researchers have proposed improvements to MSER without a drastic increase of performance. An MSER color extension, *Maximally Stable Color Region*, outperforms both an MSER-per-color-channel combination and a color blob detector [9]. Improving the MSER region distinctiveness by morphological dilation on the detected Canny edges is proposed in [10]. The improved detector shows better performance for bag of words classification, but evaluation of repeatability is not reported. MSER has been extended to *Maximally Stable Volumes* for successful segmentation of 3D medical images and paper fiber networks [11].

Fig. 3. Binary salient regions detection. Color coding: holes- blue, islands- yellow, indentations - green, protrusions- red.

Although crucial for the development of detectors, there is a shortage of evaluation benchmarks, especially for performance analysis independently of the image content. The standard *Oxford dataset* is very small: eight test sequences containing six (one base and five transformed) images of the same scene each. Every pair (base, transformed) is related via a given transformation matrix (homography) [6]. The *Freiburg dataset* contains 416 higher resolution images, generated by transforming 16 base images in order to de-tangle transformations from content [12]. The *TNT dataset* contains versions of the same viewpoint sequences with increasing resolution from 1.5 to 8 MPixel per image. Highly accurate image pair homographies are given. It is suitable for evaluating robustness to resolution rather than to transformations [7].

This paper contributes to solving the identified problems. A new regions detector, *Data-driven Morphology Salient Regions (DMRS)* is proposed and made available as open source [13]. It is related to the *Morphology-based Stable Salient Regions (MSSR)* detector that we developed in the context of humpback whale identification [14, 15]. DMRS includes a binarization, that yields a much smaller number of regions and is more stable across transformations. It has similar or higher (lighting, blur and increased resolution) repeatability compared to MSER, while detecting perceptually salient regions (Fig. 1). Figure 2 (c), (d) illustrates that DMSR finds salient regions repeatedly on the leatherback turtle images, unlike MSER. Also, we made an openly available dataset, OxFrei, combining the natural homographies of the Oxford and the higher resolution images of the Freiburg datasets [13].

2. DATA-DRIVEN MORPHOLOGY SALIENT REGIONS DETECTION

MSER and MSSR decompose a gray-scale image into binary cross-sections and evaluate the stability of the connected components (CCs) or accumulate saliency masks. DMSR starts with a data-driven binarization, producing one binary image, thus transforming the problem into binary saliency.

2.1. Binary Salient Regions Detection

We claim that the perceptual saliency in a binary image of a structured scene $\mathbf{B} : \mathcal{D} \subset \mathcal{Z}^2 \rightarrow \{0, 1\}$ (1-white, 0-black) is only due to the spatial layout of the image regions [15]. There

ISS	A CC $S^i_{fb} = \{\mathbf{p} \in \mathcal{D}, \forall \mathbf{p} = foreground,$ $\forall \mathbf{q} \in \partial S^i_{fb}, \mathbf{q} = background, \mathbf{q} \notin \partial \mathbf{B}\}$,
2 types	S^i_{10} (islands), S^i_{01} (holes); $\mathbf{S}^i = S^i_{01} \cup S^i_{10}$
BSS	$S^b_{fb} : \{\mathbf{p} \in S^b_{fb} \subset \mathcal{B}^f, \forall \mathbf{p} = foreground,$ $\mathbf{q} \in \partial S^b_{fb} \subset \partial \mathcal{B}^f, \forall \mathbf{q} = background\}$, $\lvert\partial\mathcal{B}^f\rvert - \lvert\partial(\mathcal{B}^f \backslash S^b_{fb})\rvert < 2\pi r$
2 types	S^b_{10} (protr.), S^b_{01} (indent.); $\mathbf{S}^b = S^b_{01} \cup S^b_{10}$
Regions	$\mathbf{S} = \mathbf{S}^i$ (DMSR); $\mathbf{S} = \mathbf{S}^i \cup \mathbf{S}^b$ (DMSRA)

Table 1. Binary saliency definitions used in Section 2.1.

are 4 types of salient regions. The 2 types of *inner salient structures (ISS)* are (1) *holes* – set of connected black pixels entirely surrounded by white pixels, and (2) *islands* – set of connected white pixels surrounded by black ones, i.e. inverse of holes. A significant connected component $\mathcal{B}^1$ is defined as a CC with area proportional to the image area by Λ. The radius of the morphological structuring element is r and the area opening parameter for noise filtering is λ. The 2 *boundary salient structures (BSS)* are (3) *protrusions*- set of white pixels on the border of a significant CC, which if pinched off from the CC, will increase its boundary with no more than $2\pi r$, and (4) the *indentations*- protrusions inverse.

The types are also valid for the MSSR detector. The regions are obtained from $\mathbf{B}$ by morphological operations: hole filling, top hat and area opening, for details see [14, 15]. The ISS are similar to the definition of the MSER+ and MSER- regions [1]. In this paper, detectors using only ISS, i.e., directly comparable to MSER, are denoted by DMSR/MSSR, while DMSRA/MSSRA are detectors using all region types. These definitions are summarized in Table 1 and the exact shaped regions from a synthetic 100×100 binary image with parameters $\Lambda = 100$, $r = 5$ and $\lambda = 10$ are shown on Fig. 3.

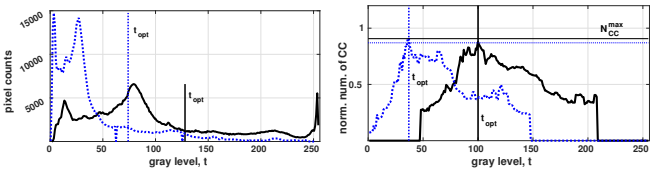
2.2. Data-driven binarization

Any gray-scale image $\mathbf{I} : \mathcal{D} \subset \mathcal{Z}^2 \rightarrow \mathcal{T}$, where $\mathcal{T} = \{0, 1, \ldots, t_{max}\}$ and $t_{max} = 2^n - 1 = 255$ is the maximum gray value encoded by $n = 8$ bits, can be decomposed into cross-sections at every possible level t: $\mathbf{I} = \sum_{t \in \mathcal{T}} CS_t(\mathbf{I})$. Obtaining a section at level t is equivalent to thresholding the image at threshold t: $CS_t(\mathbf{I}) = 1.(\mathbf{I} > t) + 0.(\mathbf{I} < t)$ is a binary image. Three sets of connected components in $CS_t(\mathbf{I})$ are defined: $\mathcal{A}_t$- *all*, $\mathcal{L}_t$- the *large* and $\mathcal{V}_t$- the *very large* CCs. The size of each CC category is defined by $\Lambda_{\mathcal{L}}$ and $\Lambda_{\mathcal{V}}$ fraction of the image area $A_{\mathbf{I}}$. Let us denote the normalized number of elements in a set ($|\cdot|$) by $\|\cdot\| = |\cdot| / \max_{t \in \mathcal{T}} |\cdot|$. Finding the optimal threshold t_{opt} is then defined as:

$$t_{opt} = \arg\max_{t \in \mathcal{T}} (w^{\mathcal{A}} \|\mathcal{A}_t\| + w^{\mathcal{L}} \|\mathcal{L}_t\| + w^{\mathcal{V}} \|\mathcal{V}_t\|),$$

where $w^{\cdot}$ are the weights per set of CCs.

The standard Otsu thresholding does not select a single stable CS, while choosing t_{opt} ensures stable number of regions across transformations (see Figs. 4 and 5 for lighting).



(a) Otsu (b) Max number CC

Fig. 4. Finding the optimal threshold for two images from the 'Leuven' sequence (Oxford dataset, lighting): the base image- solid black line, the forth image - dotted blue line.

(a) (b) (c)

Fig. 5. Binarization of two images of the 'Leuven' sequence (lighting). Top row- base image, bottom row- forth image; (a) gray scale; (b) Otsu binarization, (c) proposed binarization.

After the data-driven binarization, the DMSR detector finds the set of affine-covariant regions $\mathbf{S}$ from the single binary image $CS_{t_{opt}}$ as described in Section 2.1 and [14, 15]. DMSR produces fewer non-overlapping and perceptually salient regions (visualized by their equivalent ellipses, not exact shapes) compared to MSER (see Figures 6 and 7).

3. PERFORMANCE EVALUATION

The *repeatability score* (R) and the *number of correspondences* (N_C) are the main performance evaluation measures [6]. The maximum overlap error between matching regions is 40%. The R score between a pair of base and transformed image, $(\mathbf{I}_B, \mathbf{I}_T)$, is the ratio between N_C in the common image part and the smaller number of regions in the pair. The structured scenes from each dataset are considered and five detectors are evaluated: MSER, MSSR(A) and DMSR(A). The MSER software is used with its default settings and the (D)MSSR(A) parameters are: $r = 0.02 * \sqrt{A_{\mathbf{I}}/\pi}$, $\lambda = 3r$, $\Lambda_{\mathcal{L}} = 0.001$, $\Lambda_{\mathcal{V}} = 0.01$ and $w = 0.33$. All performance plots and detected regions on all data are available online [13].

3.1. Oxford dataset

Each image sequence of the Oxford dataset consists of one base and five increasingly distorted images [6]. They are obtained independently of each other and the homographies between each pair $(\mathbf{I}_B, \mathbf{I}_T)$ are the provided ground truth. Each

Fig. 6. Region detectors on the base image of the 'Graffiti' sequence, Oxford dataset. Left: MSER, right: DMSR

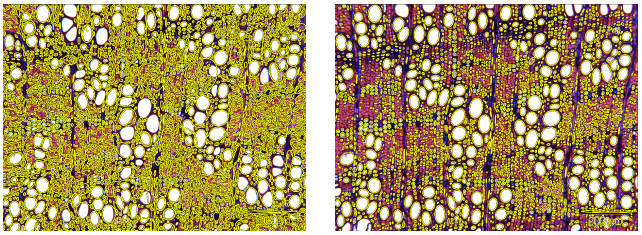


Fig. 7. Salient region detectors on microscopy wood images. Left: MSER (every second region is shown), right: DMSR

sequence can be used to test only one transformation T. For the viewpoint ('Graffiti'), the best R is achieved by DMSR with up to 72%, which is 10% more than the second performing MSER for a $40°$ change. DMSR is performing worse on the scale ('Boat'), but as good as or better than MSER on the blur ('Bikes', Fig. 8) and lighting ('Leuven') sequences.

3.2. OxFrei dataset

The creators of the Freiburg dataset separated T from the image content by applying a few transformations (alas not fully documented) to different base images [12]. To address the shortage of evaluation datasets, we release OxFrei - a combination of the strong features of the Oxford and Freiburg datasets [13]. We transform the Freiburg base images with all homographies of the Oxford dataset. In this way, we create 54 images in 9 structured scenes each under realistic blur, lighting, scaling and viewpoint transformations.

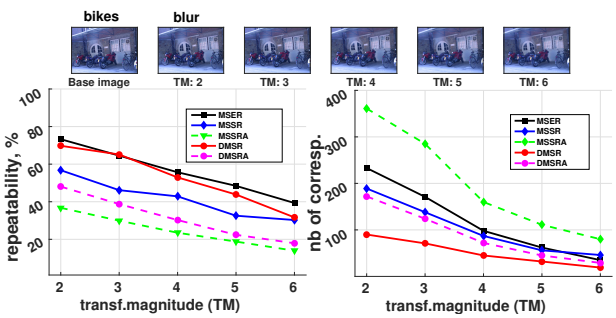


Fig. 8. Region detection on 'Bikes', Oxford dataset.

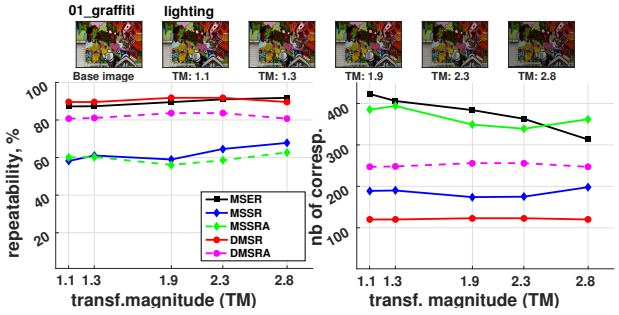


Fig. 9. Region detectors on '01_graffiti', OxFrei dataset.

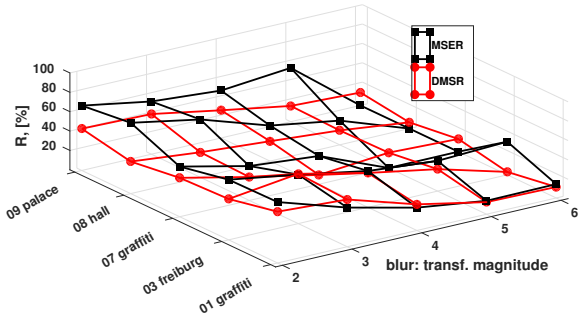


Fig. 10. Robustness of region detectors to blur on five sequences of the OxFrei dataset.

The dataset allows a transformation independent robustness study by comparing performance on all data subject to the same realistic T. Figures 9 and 10 show R for lighting for one and blur for a few sequences. The standard plots (like on Fig. 9) are cross-sections along the data dimension of the 3D plots (like on Fig. 10) for one sequence. The 2D plots for all OxFrei experiments are online [13]. We conclude that MSER is better for zoom and viewpoint (the latter contradicting with the result on the single Oxford 'Graffiti' sequence), while DMSR is robust to lighting and blur (Fig. 10).

3.3. TNT hi-res benchmark

The R score of all detectors from the Oxford study drops down on hi-res images [7]. On the 'underground' sequence from the TNT set, MSER loses up to 25% between 1.5 Mpx (R_1) and 8Mpx (R_4) resolutions. On the contrary, DMSR increases the R score as resolution increases on 'underground' and 'posters' sequences with up to 75%, which is a 15% increase from R_1 to R_4 for a 40° viewpoint change (Figure 11).

3.4. Animal and plant biometrics

The DMSR detector has been compared to MSER on several small animal individual photo-ID datasets (humpback whales, leatherback turtles, newts) and on a wood species identification dataset [15, 16, 17]. In all cases DMSR produces fewer

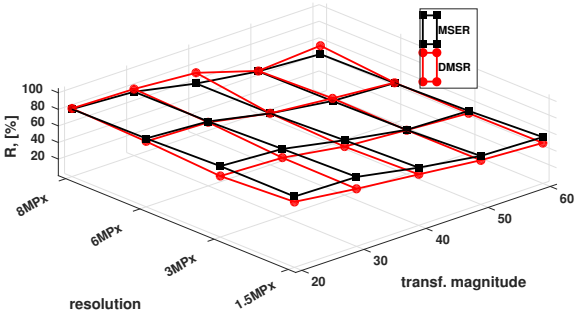


Fig. 11. Robustness of region detectors to image resolution and viewpoint. 'Posters', TNT dataset.

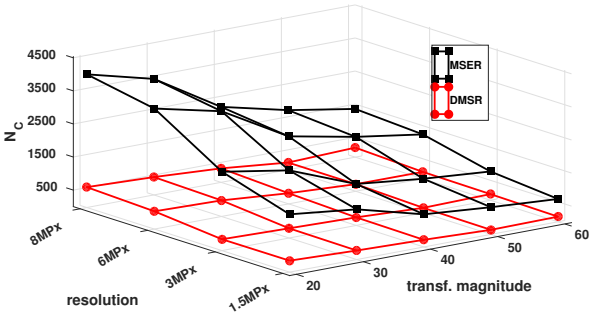


Fig. 12. Number of region correspondences versus image resolution and viewpoint. 'Posters', TNT dataset.

and perceptually more acurate salient regions as illustrated by Figures 1, 2 and 7. For the wood microscopy images, it is not possible to obtain accurate statistics on the cell properties with MSER, while the detected DMSR regions enable such wood anatomy research.

On all datasets, the N_C plane of DMSR has the lowest values and the least slope of all detectors (Figs. 8, 9, right and Fig. 12) [13]. The number of detected DMSR regions is up to an order of magnitude lower compared to MSER - crucial for the efficiency of the following matching step on large-scale image data. Using all 4 types of regions does not improve performance over using only ISS (holes and islands) with the exception of detecting markings on humpback whale tails.

4. CONCLUSIONS

Combining data-driven binarization with morphological operations yeilds a region detector with comparable to superior performance to MSER on various datasets. DMSR produces a much smaller number of regions- a very desired property in large scale processing. It copes better with blur, lighting and increased resolution and detects perceptually salient regions, which makes it a good choice for scientific imagery analytics. For detection evaluation, high-resolution transformation-independent datasets should become the standard.

5. REFERENCES

[1] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," in *Proceedings BMVC*, 2002, pp. 36.1–36.10.

[2] Foncubierta-Rodrguez et al., "Region-based volumetric medical image retrieval," in *SPIE Medical Imaging: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, 2013.

[3] S. Escalera, P. Radeva, and O. Pujol, "Complex salient regions for computer vision problems," in *CVPR*, 2007.

[4] H. Kuehl and T. Burghardt, "Animal Biometrics: quantifying and detecting phenotypic appearance," *Trends in Ecology & Evolution*, vol. 28, pp. 432–441, 2013.

[5] N. *et al.*, Kumar, "Leafsnap: Computer Vision System for Automatic Plant Species Identification," in *The 12th European Conference on Computer Vision (ECCV)*, October 2012, pp. 502–516.

[6] K. Mikolajczyk et al., "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, November 2005.

[7] K. Cordes, B. Rosenhahn, and J. Ostermann, "High-Resolution Feature Evaluation Benchmark," in *The 15th International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2013, pp. 327–334.

[8] R. Kimmel et al., "Are MSER Features Really Interesting?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2316–2320, 2011.

[9] Per-Erik Forssén, "Maximally Stable Color Regions for Recognition and Matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[10] Sh. Wang et al., "Enhanced Maximally Stable Extremal Regions with Canny Detector and Application in Image Classification," *Journal of Computational Information Systems*, vol. 10, no. 14, pp. 6093–6100, 2014.

[11] M. Donoser and H. Bischof, "3D Segmentation by Maximally Stable Volumes (MSVs).," in *ICPR (1)*, 2006, pp. 63–66.

[12] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT," *CoRR*, vol. abs/1405.5769, 2014.

[13] E. Ranguelova, "Large scale imaging: Data, software, results," DOI: http://dx.doi.org/10.5281/zenodo.45156, Jan. 2016.

[14] E. B. Ranguelova and E. J. Pauwels, "Morphology-Based Stable Salient Regions Detector," in *Proceedings of International Conference on Image and Vision Computing New Zealand*, 2006, pp. 97 – 102.

[15] E. B. Ranguelova and E. J. Pauwels, "Saliency Detection and Matching for Photo-Identification of Humpback Whales," *International Journal on Graphics, Vision and Image Processing*, 2006.

[16] E. Pauwels, P. de Zeeuw, and D. Bounantony, "Leatherbacks matching by automated image recognition," in *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, 8th Industrial Conference, ICDM 2008, Leipzig, Germany, July 16-18, 2008, Proceedings*, 2008, pp. 417–425.

[17] Images courtesy to Frederic Lens, "Microscopy images wood," http://www.naturalis.nl/en/, Naturalis Biodiversity Center, Leiden, The Netherlands.