

# Shape- and Moment- Invariant Descriptor for Structured Images

Anonymous Submission

*Anonymous Affiliation*

## Abstract

Finding correspondences between two images to determine if they depict the same scene is a fundamental, yet challenging task. To cope with different viewpoints and lighting conditions, local salient regions are detected invariantly to transformations and encoded by descriptors such as Scale-invariant Feature Transform (SIFT) or Speeded-up Robust Features (SURF). While using the image intensities around the centroid of the region to compute a SIFT-type descriptor often works well, we argue that for structured scenes it is beneficial to use descriptors based on the shape of the regions. We propose a 20-dimensional Shape- and Moment- Invariant (SMI) descriptor and show that it outperforms the 64-dimensional SURF on two benchmark datasets in terms of precision with similar or higher accuracy, while having a better scalability.

**Keywords:** image matching, affine-invariant descriptor, shape invariants, moment invariants

## 1 Introduction

Automatically determining whether two images depict partially the same physical scene is a fundamental computer vision problem such as baseline stereo matching, image retrieval, etc. [Escalera et al., 2007, Matas et al., 2002]). The approach is to *detect* local (to cope with partial overlap) features, followed by matching of their *descriptors*. A class of such features are local regions, corresponding to the same image patches, detected independently in each of the two images. Many detectors and descriptors are invariant to photometric (due to different sensors and lighting) and affine geometric transformations (due to different viewpoints). In recent years, the approach of using large datasets of image patch correspondences has been established, [Snavely et al., 2008, Zagoruyko and Komodakis, 2015]. However, deep learning is not applicable when the *structured* images with homogeneous regions with distinctive boundaries, are only few. Such is the case, for example, in some scientific applications [Ranguelova, 2016].

The Maximally Stable Extremal Regions (MSER) detector has become the standard in computer vision [Matas et al., 2002]. It is often used in combination with a histogram-of-gradients type descriptor such as Scale-invariant Feature Transform (SIFT) or Speeded-Up Robust Features (SURF) [Bay et al., 2008], computed from image intensities around the centroids of the MSER regions. We argue that using the shape and moment information of the regions encoded by a *Shape- and Moment- Invariant (SMI)* descriptor is beneficial, compared to image intensities around the region's centroid. Figure 1 illustrates two cases of image pairs, one depicting the same scene and the other not, where SMI outperforms SURF applied on pre-detected MSER regions.

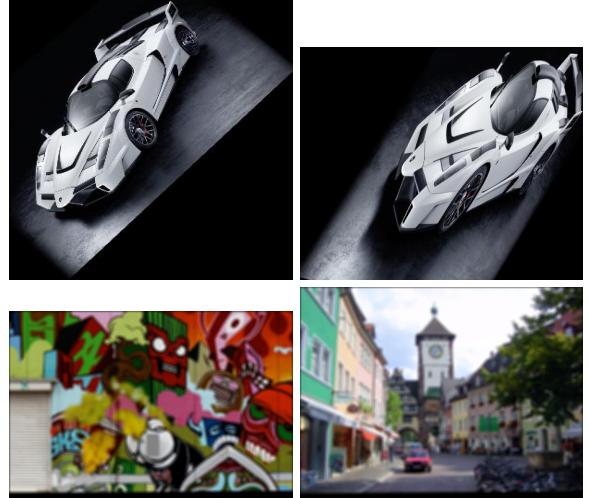


Figure 1: “Is it the same object or scene?” Matching two images under different transformations using MSER regions. *Top image pair* (scale and viewpoint): SURF descriptor yields false negative (similarity score 0.096), while SMI - true positive (0.89). *Bottom image pair* (blur): SURF gives false positive (0.27), while SMI true negative (-0.11).

## 2 Related work

The literature describes large number of local detectors and descriptors. For a recent introduction and overview the reader is referred to [Hassaballah et al., 2016]. Here, we mention very briefly only the closely related work.

A comparative performance evaluation of number of detectors has concluded that the MSER is the best performing region detector for structured scenes [Mikolajczyk et al., 2005]. Since then, MSER has been integrated in MATLAB, OpenCV, VLFeat, etc., making it the default baseline detector. However, despite its success, the detector has several drawbacks, which have been addressed by improved detectors, including the Data-driven Morphology Salient Regions detector (DMSR) [Ranguelova, 2016]. Here, we propose to use a Binary detector (BIN) using the first step of DMSR construction: data-driven binarization explained in [Ranguelova, 2016], with either all regions or only regions with large area ( $A_{region} \geq f_A \cdot A_{Image}$ ).

Another comparative performance evaluation of number of region descriptors have concluded that the "region-based SIFT descriptor" is the best performing for structured images [Mikolajczyk and Schmid, 2005]. Since we are interested in describing the shape of the detected regions, we have chosen efficient shape descriptors, known as moment invariants. Flusser et al. pointed our the dependency in the early set of 7 Hu moments and developed a coherent theory and general framework for derivation of Affine Moment Invariants (AMIs) using graph representation [Suk and Flusser, 2004, Flusser et al., 2009].

Research has been performed not only to determine the best region detector and descriptor, but also the best detector - descriptor combination. For example, the conclusion of the experiments in [Dahl et al., 2011] is that the best combination is DOG or MSER detector and SIFT (SURF was not included in the experiments) or DAISY descriptors. SURF has been introduced as an improvement over SIFT and since has become a standard descriptor in many computer vision software libraries, making it the default baseline descriptor [Bay et al., 2008]. Hence, we have chosen MSER - SURF as the baseline detector - descriptor combination.

## 3 Image matching with Shape- and Moment- Invariant descriptor

We propose a set of several Shape and Moment Invariants (SMI) derived from the binary shapes of the detected regions as a region descriptor. The SMI descriptor contains *shape invariants* and *moment invariants*.

**Shape invariants.** A shape of a region  $R_i$  can be described by a set of simple properties either of the original shape or of the equivalent (up to the second order moments) ellipse  $E_i$ . These are: the region's area  $a_i$ , the area  $a_i^c$  of the region's convex hull, the length  $\mu_i$  of the major and  $\nu_i$  of the minor axes of  $E_i$  and the distance  $\phi_i$  between the foci of the ellipse. From these properties, a set of shape affine invariants are defined in Table 1.

Invariant	Definition	Description
Relative Area	$\tilde{a}_i = a_i / A$	region's area normalized by the image area $A$
Ratio Axes Lengths	$r_i = \nu_i / \mu_i$	ratio between $E_i$ minor and major axes lengths
Eccentricity	$e_i = \phi_i / \mu_i$	$e_i \in [0, 1]$ (0 is a circle, 1 is a line segment.)
Solidity	$s_i = a_i / a_i^c$	proportion of the convex hull pixels, that are also in the region.

Table 1: Simple shape invariants.

**Affine Moment Invariants.** If  $f(x, y)$  is a real-valued image with  $N$  points, the AMI functional is defined by

$$M(f) = \int_{-\infty}^{\infty} \prod_{k,j=1}^N C_{kj}^{n_{kj}} \cdot \prod_{l=1}^N I(x_l, y_l) dx_l dy_l, \quad (1)$$

where  $n_{kj}$  are non-negative integers and  $C_{kj} = x_k y_j - x_j y_k$  is the cross-product (graph edge) of points (nodes)  $(x_k, y_k)$  and  $(x_j, y_j)$ , [Suk and Flusser, 2004]. For full details of the AMI's theory the reader is referred to [Flusser et al., 2009]. We use the set of 16 irreducible AMIs of  $N = 4$ th order, which are the functional coefficients  $\{m_{ij}, j = 1 \dots 16\}$ , as implemented by the authors in an open source MATLAB software.

Hence, the final descriptor for the  $i$ -th region is a 20 element feature vector  $SMI_i = (\tilde{a}_i, r_i, e_i, s_i, m_{i1}, \dots, m_{i16})$ .

**Matching.** Lets  $SMI1$  and  $SMI2$  be  $n1 \times 20$  and  $n2 \times 20$  matrices, where each row is the SMI descriptor for the  $n1$  and  $n2$  regions detected via MSER or BIN (all/largest) detector in the pair of images  $\langle I1, I2 \rangle$ . We compare exhaustively  $SMI1$  and  $SMI2$  with Sum of square differences metric. The matching threshold for selection of the strongest matches is  $mt$ , the max ratio threshold for rejecting ambiguous matches is  $mr$ , the confidence of a match is  $mc$  and only unique matches are allowed. Then, we select the top quality matches above a cost threshold  $ct$ . From those, we estimate in  $it$  iterations the affine transformation  $\tilde{T}$  between the two sets of points-centroids of the matching regions sets as average of  $nr$  runs with allowed max point distance  $md$ . The two images are then transformed  $J2 = \tilde{T}(I1)$ ,  $J1 = \tilde{T}^{-1}(I2)$  and a correlation ( $cor[X, Y] = cov[X, Y]/\sqrt{var[X]var[Y]}$ ) between the original and transformed images is used for confirmation of a true match. If the average correlation similarity between both images and their transformed versions ( $cor[I1, J1] + cor[I2, J2])/2$  is above a similarity threshold  $st$ , we declare the image pair  $\langle I1, I2 \rangle$  to be depicting (partially) the same scene.

Figure 2 illustrates the major steps of the image matching using BIN - SMI in case of viewpoint distortion. Note the better alignment in the right part of the images due to the larger number of correct matches there. The steps are the same when using MSER instead of the data-driven binarization or SURF instead of SMI descriptor.



Figure 2: Matching two images of the same scene under viewpoint transformation using BIN regions and SMI descriptor. *First column:* original images  $I1, I2$ ; *second column:* binarization; *third column:* matched BIN regions using SMI descriptor (blend view with pseudocolours); *fourth column:* overlay of the original and transformed images  $(I1, J1), (I2, J2)$ .

## 4 Performance Evaluation

We have tested the performance of the MSER and BIN (all regions and only the largest) detectors in combinations with the SURF and SMI descriptors on two datasets: Oxford (VGG) [Mikolajczyk et al., 2005] and OxFrei, [Ranguelova, 2016]. Each of the 4 structured image sequences of the Oxford set consists of 1 base and 5 increasingly distorted images. Each sequence can be used to test only one transformation  $T$ : viewpoint, scaling + rotation, decreased lighting and blur. OxFrei dataset overcomes the limitation: 9 structured scenes each with 21 images (original + 5 images for  $4T$ ) using Oxford’s real homographies. We compared all possible image pairs and assigned a flag *True/False* if a pair is depicting the same scene (see Section 3). The values of the parameters used in the experiments have been determined empirically:  $mt = mr = 1$ ,  $f_A = 2e-3$  (for BIN largest),  $it = 1000$ ,  $nr = 10$ ,  $mc = 95$ ,  $md = 8px$ ,  $ct = 0.025$ ,  $st = 0.25$ . Figure 1 illustrates the result for 2 pairs and Figure 3 for all pairs (a pixel represents image pair and a square block a sequence) of the OxFrei dataset. Note the lower number of false positives and correlation similarity variance when using the SMI descriptor.

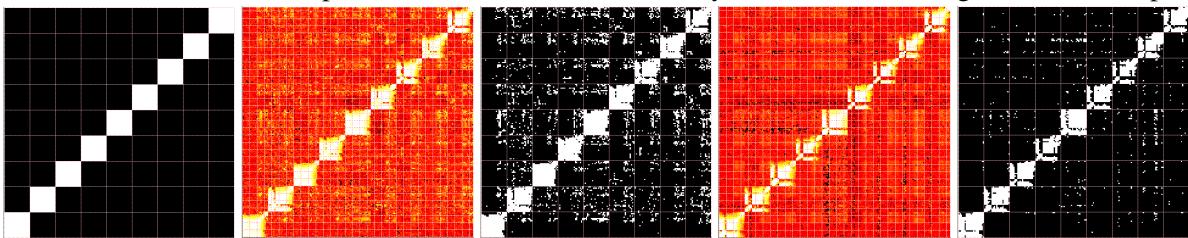


Figure 3: OxFrei dataset, MSER regions. “Is the image pair from the same scene?”: *True* = white, *False* = black. *First:* ground truth, *third:* SURF, *fifth:* SMI. Correlation similarity: the lighter, the higher. *Second:* SURF, *fourth:* SMI.

Table 2 summarizes the performance of the combinations of detectors and descriptors for the 2 datasets. When using the default MSER detector, it seems beneficial to combine with an SMI instead of the standard SURF descriptor as in both datasets almost all performance measures are improved. The BIN (all) detector does not outperform MSER in the matching task and using only the largest regions in BIN (largest) improves the recall at the expense of lower precision. The SMI descriptor achieves better precision in comparison to SURF independent of the detector or dataset. The best detector - descriptor combination is MSER - SMI when all measure are required to be high, especially accuracy and precision.

Dataset	Oxford			OxFrei		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Detector - descriptor						
MSER - SURF	<b>0.97</b>	0.97	<b>0.89</b>	0.90	0.53	<b>0.83</b>
MSER - SMI	<b>0.96</b>	<b>0.98</b>	<b>0.85</b>	<b>0.95</b>	<b>0.83</b>	<b>0.74</b>
BIN (All) - SURF	0.95	0.95	0.85	0.85	0.41	0.63
BIN (All) - SMI	0.89	<b>1</b>	0.58	0.91	0.73	0.32
BIN (Largest) - SMI	0.93	0.93	0.77	0.85	0.38	0.52

Table 2: Performance of salient region detectors and descriptors on the Oxford and OxFrei datasets.

The developed MATLAB software will be released as open source with the final paper.

## 5 Conclusion

It is not possible to use deep learning of image patches approach when trying to automatically determine if two images depict (partially) the same scene if only a few images are available. It is beneficial not to discard the shape of the salient regions detected by the detector. For structured scenes, a descriptor based on the properties of the binary regions alone performs better than one based on image intensities. The proposed shape and moment invariant, SMI, descriptor is a good choice when false positives should be minimal. In combination with the MSER detector, SMI achieves the highest precision and good accuracy and recall. In the future, the matching performance of the SMI descriptor should be tested on larger datasets.

## References

- [Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359.
- [Dahl et al., 2011] Dahl, A. L., AanÅçs, H., and Pedersen, K. S. (2011). Finding the best feature detector-descriptor combination. In *3DIMPVT*, pages 318–325. IEEE Computer Society.
- [Escalera et al., 2007] Escalera, S., Radeva, P., and Pujol, O. (2007). Complex salient regions for computer vision problems. In *CVPR*.
- [Flusser et al., 2009] Flusser, J., Suk, T., and Zitova, B. (2009). *Moments and Moment Invariants in Pattern Recognition*. Wiley.
- [Hassaballah et al., 2016] Hassaballah, M., Abdelmgeid, A. A., and Alshazly, H. A. (2016). Image Features Detection, Description and Matching. In Awad, A. and Hassaballah, M., editors, *Image Feature Detectors and Descriptors: Foundations and Applications*, volume 630 of *Studies in Computational Intelligence*, pages 11–45. Springer.
- [Matas et al., 2002] Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proceedings BMVC*, pages 36.1–36.10.
- [Mikolajczyk et al., 2005] Mikolajczyk, K. et al. (2005). A comparison of affine region detectors. *Int. J. of CV*, 65(1-2):43–72.
- [Mikolajczyk and Schmid, 2005] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630.
- [Rangelova, 2016] Rangelova, E. (2016). A Salient Region Detector for Structured Images. In *Proceedings of IEEE/ACS 13th Int. Conf. of Computer Systems and Applications (AICCSA)*, pages 1–8.
- [Snavely et al., 2008] Snavely, N., Seitz, S. M., and Szeliski, R. (2008). Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2):189–210.
- [Suk and Flusser, 2004] Suk, T. and Flusser, J. (2004). Graph method for generating affine moment invariants. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, pages 192–195.
- [Zagoruyko and Komodakis, 2015] Zagoruyko, S. and Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. *CoRR*, abs/1504.03641.