Maximally stable local description for scale selection

Gyuri Dorkó, Cordelia Schmid

Maximally Stable Local Description for Scale Selection

Gyuri Dorkó and Cordelia Schmid

INRIA Rhône-Alpes, 655 Avenue de l'Europe, 38334 Montbonnot, France
{gyuri.dorko,cordelia.schmid}@inrialpes.fr

Abstract. Scale and affine-invariant local features have shown excellent performance in image matching, object and texture recognition. This paper optimizes keypoint detection to achieve stable local descriptors, and therefore, an improved image representation. The technique performs scale selection based on a region descriptor, here SIFT, and chooses regions for which this descriptor is maximally stable. Maximal stability is obtained, when the difference between descriptors extracted for consecutive scales reaches a minimum. This scale selection technique is applied to multi-scale Harris and Laplacian points. Affine invariance is achieved by an integrated affine adaptation process based on the second moment matrix. An experimental evaluation compares our detectors to Harris-Laplace and the Laplacian in the context of image matching as well as of category and texture classification. The comparison shows the improved performance of our detector.

1 Introduction

Local photometric descriptors computed at keypoints have demonstrated excellent results in many vision applications, including object recognition [1, 2], image matching [3], and sparse texture representation [4]. Recent work has concentrated on making these descriptors invariant to image transformations. This requires the construction of invariant image regions which are then used as support regions to compute invariant descriptors. In most cases a detected region is described by an independently chosen descriptor. It would, however, be advantageous to use a description adapted to the region. For example, for blob-like detectors which extract regions surrounded by edges, a natural choice would be a descriptor based on edges. However, adapted representations may not provide enough discriminative information, and consequently, a general descriptor, such as SIFT [5], could be a better choice. Many times this leads to better performance, yet less stable representations: small changes in scale or location can alter the descriptor significantly. We found that the most unstable component of keypoint-based scale-invariant detectors is the scale selection. We have, therefore, developed a detector which uses the descriptor to select the characteristic scales. Our feature detection approach consists of two steps. We first extract interest points at multiple scales to determine informative and repeatable locations. We then select the characteristic scale for each location by identifying maximally stable local

descriptions. The chosen local description can be any measure computed on a pixel neighborhood, such as color histograms, steerable filters, or wavelets. For our experiments we use the Scale Invariant Feature Transform (SIFT) [5], which has shown excellent performance for object representation and image matching [6]. The SIFT descriptor is computed on a 4x4 grid with an 8-bin orientation histogram for each cell, resulting in a 128-dimensional vector for a given local region.

Our method for scale-invariant keypoint detection and image representation has the following properties:

- Our scale selection method guarantees more stable descriptors than state-of-the-art techniques by explicitly using descriptors during keypoint detection. The stability criteria is developed to minimize the variation of the descriptor for small changes in scale.
- Repeatable locations are provided by interest point detectors (e.g. Harris), and therefore they have rich and salient neighborhoods. This consequently helps to choose repeatable and characteristic scales. We verify this experimentally, and show that our selection competes favorably with the best available detectors.
- The detector takes advantage of the properties of the local descriptor. This can include invariance to illumination or rotation as well as robustness to noise. Our experiments show that the local invariant image representation extracted by our algorithm leads to significant improvement for object and texture recognition.

Related Work. Many different scale- and affine-invariant detectors exist in the literature. Harris-Laplace [7] detects multi-scale keypoint locations with the Harris detector [8] and the characteristic scales are determined by the Laplacian operator. Locations based on Harris points are very accurate. However, scale estimation is often unstable on corner-like structures, because it depends on the exact corner location, i.e., shifts by one pixel may modify the selected scale significantly. The scale-invariant Laplacian detector [9] selects extremal values in location-scale space and finds blob-like structures. Blobs are well localized structures, but due to their homogeneity, the information content is often poor in the center of the region. The detector of Kadir et al. [10] extracts circular or elliptical regions in the image as maxima of the entropy scale-space of region intensity histograms. It extracts also blob-like structures, and has shown to be a more robust representation for some object categories [10]. Mikolajczyk et al. [11] show that it performs poorly for image matching, which might be due to the sparsity of the scale quantization. Edge and structure based scale-invariant detectors [12–14] also exist in the literature. Some of them have been evaluated in [11] and apart from MSER [14] have shown to be inferior to Harris-Laplace or Hessian-Laplace. The MSER (Maximally Stable Extremal Regions) detector [14] defines extremal regions as image segments where each inner-pixel intensity value is less (greater) than a certain threshold, and all intensities around the boundary are greater (less) than the same threshold. An extremal region is *maximally stable* when the area (or the boundary length) of the segment changes the least with

respect to the threshold. This detector works particularly well on images with well defined edges, but it is less robust to noise and is not adapted to texture-like structures. It usually selects fewer regions than the other detectors.

Viewpoint invariance is sometimes required to achieve reliable image matching, object or texture recognition. Affine-invariant detectors [7, 9, 10, 12, 14] estimate the affine shape of the regions to allow normalization of the patch prior to descriptor computation. Lindeberg and Gårding [9] use an *affine adaptation* process based on the second moment matrix for the Laplacian detector. The affine extension of Harris-Laplace [7] is also based on this affine adaptation. The adaptation procedure is a post-processing step for the scale-invariant detections.

Overview. Our paper is organized as follows. In Section 2 we present our scale selection technique Maximally Stable Local SIFT Description (MSLSD) and introduce two detectors, Harris-MSLSD and Laplacian-MSLSD. We then compare their performance to Harris-Laplace and the Laplacian. In Section 3 we evaluate the detectors for image matching using a publicly available framework. Section 4 reports results for object category and texture classification. Finally, in Section 5 we conclude and outline future extensions.

2 Maximally Stable Local Description

In this section we present our method for selecting characteristic scales at keypoints and discuss the properties of our approach. We address two key features of interest point detectors: repeatability and description stability. *Repeatability* determines how well the detector selects the same region under various image transformations, and is important for image matching. In practice, due to noise and object variations, the corresponding regions are never exactly the same but their underlying descriptions are expected to be similar. This is what we call the *description stability*, and it is important for image matching and appearance based recognition.

The two properties, *repeatability* and *descriptor stability*, are in theory contradictory. A homogeneous region provides the most stable description, whereas its shape is in general not stable. On the other hand, if the region shape is stable, for example using edges as region boundaries, small errors in localization will often cause significant changes of the descriptor. Our solution is to apply the Maximally Stable Local Description algorithm to interest point locations only. These points have repeatable locations and informative neighborhoods. Our algorithm adjusts their scale parameters to stabilize the descriptions and rejects locations where the required stability cannot be achieved. The combination of repeatable location selection and descriptor stabilized scale selection provides a balanced solution.

Scale-invariant MSLSD detectors. To select characteristic locations with high repeatability we first detect interest points at multiple scales. We chose two widely used complementary methods, Harris [8] and the Laplacian [15, 16]. Harris detects corners, i.e., locations where the intensity varies significantly in

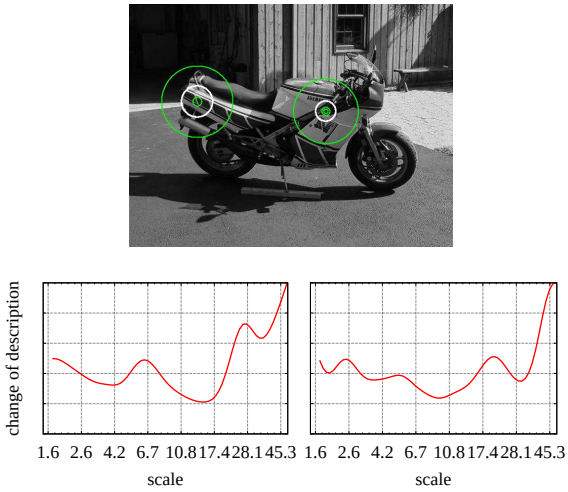

Fig. 1. Two examples for scale selection. The left and right graphs show the change in the local description as a function of scale for the left and right points respectively. The scales for which the functions have local minima are shown in the image. The bright thick circles correspond to the global minima.

several directions. The Laplacian detects blob-like structures. Its multi-scaled version detects extrema of the 2D Laplacian operator on multiple scales.

The second step of our approach selects the characteristic scales for each keypoint location. We use *description stability* as criterion for scale selection: the scale for each location is chosen such that the corresponding representation (in our case SIFT [5]) *changes the least* with respect to scale. Fig. 1 illustrates our selection method for two Harris points. The two graphs show how the descriptors change as we increase the scale (the radius of the region) for the two keypoints. To measure the difference between SIFT descriptions we use the Euclidean distance as in [5]. The minima of the functions determine the scales where the descriptions are the most stable; their corresponding regions are depicted by circles in the image. Our algorithm selects the *absolute minimum* (shown as bright thick circles) for each point. Multi-scale points which correspond to the same image structure often have the same absolute minimum, i.e. result in the same region. In this case only one of them is kept in our implementation. To limit the number of selected regions an additional threshold can be used to reject unstable keypoints, i.e., if the minimum change of description is above a certain value the keypoint location is rejected. For each point we use a percentage of the maximum change over scales at the point location, set to 50% in our experiments.

Our algorithm is in the following referred to as *Maximally Stable Local SIFT Description* (MSLSD). Depending on the location detector we add the prefix H for Harris and L for Laplacian, i.e. H-MSLSD and L-MSLSD.

Illumination and rotation invariance. Our detectors are robust to illumination changes, as our scale selection is based on the SIFT descriptor. SIFT is

normalized to unit length, and therefore offers invariance to scalar changes in image contrast. Since the descriptor is based on gradients, it is also invariant to an additive constant change in brightness, i.e., it is invariant to affine illumination changes.

The rotation invariance for SIFT can be achieved by extracting the dominant orientation and rotating the patch in this direction. If the keypoints have poorly defined orientations, the resulting descriptions are unstable and noisy. In our algorithm we orienting the patch in the dominant direction prior to the descriptor computation for each scale. Maximal description stability is then found for locations with well defined local gradients. In our experiments a *-R* suffix indicates rotation invariance. Experimental results in Section 4 show that our integrated estimation of the dominant orientation can significantly improve results.

Affine invariance. The affine extension of our detector is based on the affine adaptation in [9, 17], where the shape of the elliptical region is determined by the second moment matrix of the intensity gradient. However, unlike other detectors [4, 7], we do not use this estimation as a post-processing step after scale selection, but estimate the elliptical region prior to the descriptor computation for each scale. When the affine adaptation is unstable, i.e., sensitive to small changes of the initial scale, the descriptor changes significantly and the region is rejected. This improves the robustness of our affine-invariant representation. In our experiments an *-Aff* suffix indicates affine invariance. Full affine invariance requires rotation invariance, as the shape of each elliptical region is transformed into a circle reducing the affine ambiguity to a rotational one. Rotation normalization of the patch is, therefore, always included when affine invariance is used in our experiments.

3 Evaluation for image matching

This section evaluates the performance of our detectors for image matching based on the evaluation framework in [11], i.e., for the criteria repeatability rate and matching score. We compare our results to Harris-Laplace and LoG.

The repeatability rate measures how well the detector selects the same scene region under various image transformations. Each sequence has one reference image and five images with known homographies to the reference image. Regions are detected for the images and their accuracy is measured by the amount of overlap between the detected region and the corresponding region projected from the reference image with the known homography. Two regions are matched if their *overlap error* is sufficiently small:

$$1 - \frac{R_{\mu_a} \cap R_{(H^T \mu_b H)}}{R_{\mu_a} \cup R_{(H^T \mu_b H)}} < \epsilon_O$$

where R_μ is the elliptic or circular region extracted by the detector and H is the homography between the two images. The union $(R_{\mu_a} \cup R_{(H^T \mu_b H)})$ and the intersection $(R_{\mu_a} \cap R_{(H^T \mu_b H)})$ of the detected and projected regions are computed numerically. As in [11] the maximum possible overlap error ϵ_O is set to 40% in

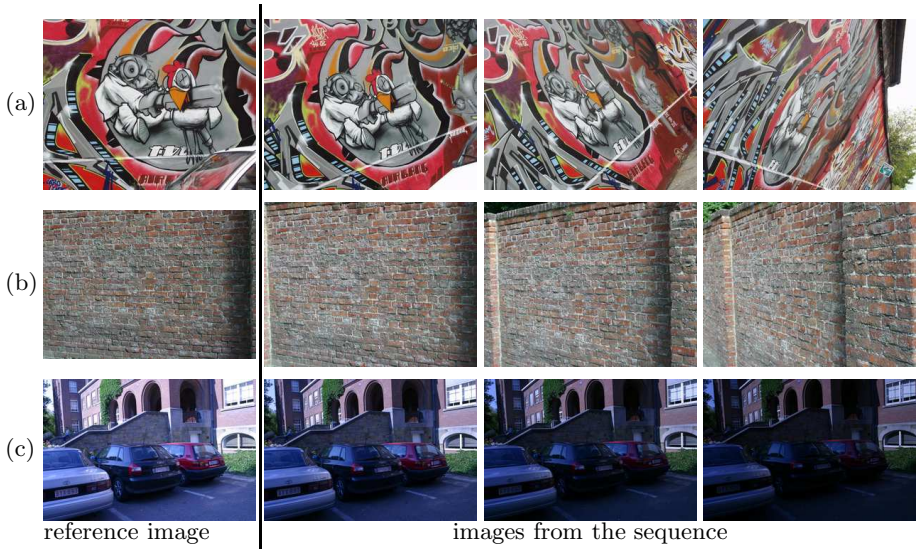

Fig. 2. Image sequences used in the matching experiments. (a), (b) Viewpoint change. (c) Illumination change. The first column shows the reference image. These sequences may be downloaded from http://www.robots.ox.ac.uk/∼vgg/research/affine/index.html.

our experiments. The *repeatability score* is the ratio between the correct matches and the smaller number of detected regions in the pair of images.

The second criterion, the matching score, measures the discriminative power of the detected regions. Each descriptor is matched to its nearest neighbor in the second image. This match is marked as correct if it corresponds to a region match with maximum overlap error 40%. The matching score is the ratio between the correct matches and the smaller number of detected regions in the pair of images.

3.1 Viewpoint changes

The performance of our detectors for viewpoint changes is evaluated on two different image sequences with viewpoint changes from 20 to 60 degrees. Fig. 2(a) shows sample images of the graffiti sequence. This sequence has well defined edges, whereas the wall sequence (Fig. 2(b)) is more texture-like.

Fig. 3 shows the repeatability rate and the matching score as well as the number of correct matches for different affine-invariant detectors. The ordering of the detectors is very similar for the criteria repeatability rate and matching score, as expected. On the graffiti sequence (Fig. 3, first row) the original Harris-Laplace (H-L-Aff) detector performs better than H-MSLSD-Aff. On the wall sequence results for H-MSLSD-Aff are slightly better than for H-L-Aff. This shows that the Laplacian scale selection provides good repeatability mainly in the presence

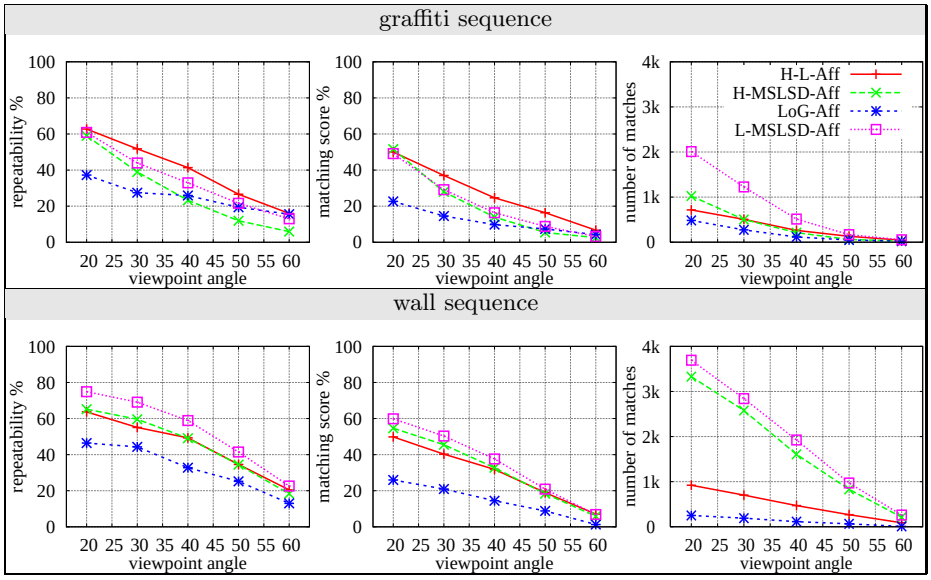

Fig. 3. Comparison of detectors for viewpoint changes. The repeatability rate, matching score and the number of correct matches are compared on the graffiti (first row) and on the wall (second row) sequence.

of well defined edges. In case of the Laplacian our detector (L-MSLSD-Aff) outperforms the original one (LoG) for both sequences. This can be explained by the fact that LoG-Aff detects a large number of unstable (poorly repeatable) regions for nearly parallel edges, see Fig. 4. A small shift or scale change of the initial regions can lead to completely different affine parameters of LoG-Aff. These regions are rejected by L-MSLSD-Aff, as the varying affine parameters cause large changes in the local description over consecutive scale parameters. Note that in case of affine divergence both detectors reject the points. This example clearly shows that description stability leads to more repeatable regions. In case of natural scenes, as for example the wall sequence, this advantage is even more apparent, i.e., the difference between L-MSLSD-Aff over LoG-Aff is higher than for the graffiti sequence.

LoG LoG-Aff L-MSLSD-Aff

Fig. 4. Output of LoG detector on part of a graffiti image: the standard LoG detector (left), affine-invariant LoG (middle) and L-MSLSD-Aff (right).

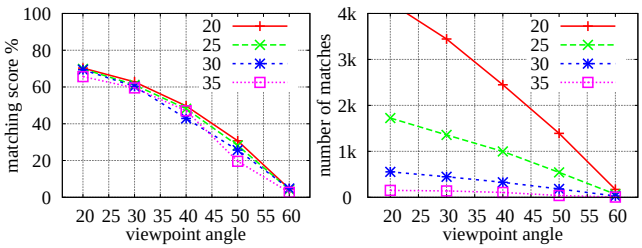

Fig. 5. Comparison of the matching score and the number of correct matches for several thresholds for the multi-scale Laplacian $(20, 25, 30, 35)$. Results are given for L-MSLSD on the wall sequence. A higher threshold results in less detections, and consequently a smaller number of absolute matches (second column).

We can observe that we obtain a significantly higher number of correct matches with our detectors. This is due to a larger number of detected regions. This could increase the probability of accidental matches. To ensure that this did not bias our results—and to evaluate the effect of the detected region density—we compared the performance for different Laplacian thresholds for the L-MSLSD detector. Note that the Laplacian threshold determines the number of detections in location space, whereas the scale threshold rejects unstable locations and remains fixed throughout the paper. Fig. 5 shows that as the number of correct matches gradually decrease, the quality of the descriptors (matching score) stays the same. Consequently, we can conclude that the quality of the detections does not depend on the density of the extracted regions.

Fig. 6 shows that in case of small viewpoint changes the scale-invariant versions of the detectors perform better that the ones with affine invariance. It also allows to compare the scale-invariant detectors. On the graffiti images the original H-L performs better that its affine adapted version until 30° of viewpoint change. For our detector this transition occurs later around 40°. In the case of L-MSLSD and LoG the curves cross around 35° and 40° respectively. On the wall sequence it is almost never helpful to use the affine adaptation, scale invariance is sufficient until $55 - 60°$. We can conclude that the use of affine invariance is not necessary unless the viewpoint changes are significant, and that it is more helpful in case of structured scenes. We can also observe that the scale-invariant versions H-L and H-MSLSD give comparable results for the graffiti sequence, whereas in the case of affine invariance H-L-Aff outperforms H-MSLSD-Aff. In the other cases, our scale-invariant detectors outperform their standard versions. In addition, the improvement of our detectors over the standard versions is more significant for scale invariance than for affine invariance, in particular for the Laplacian and the wall sequence.

3.2 Illumination changes

Experiments are carried out for the Leuven sequence (Fig. 2 (c)), i.e., images of the same scene under gradually reduced camera aperture. Fig. 7 shows that

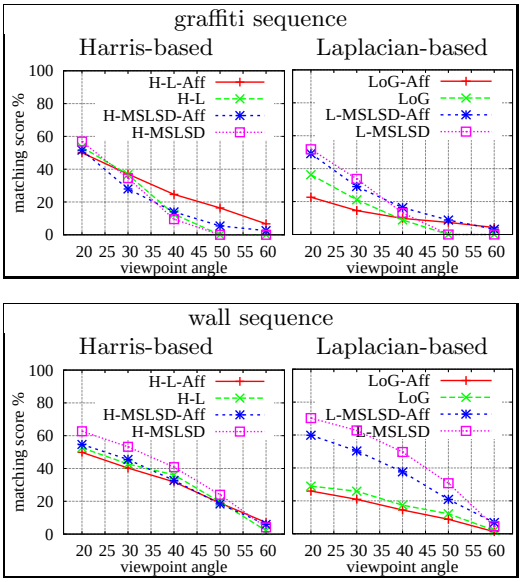

Fig. 6. Comparison of detectors with and without affine invariance on the graffiti (first row) and the wall (second row) sequence. The first column shows results for Harris- and the second for Laplacian-based detectors.

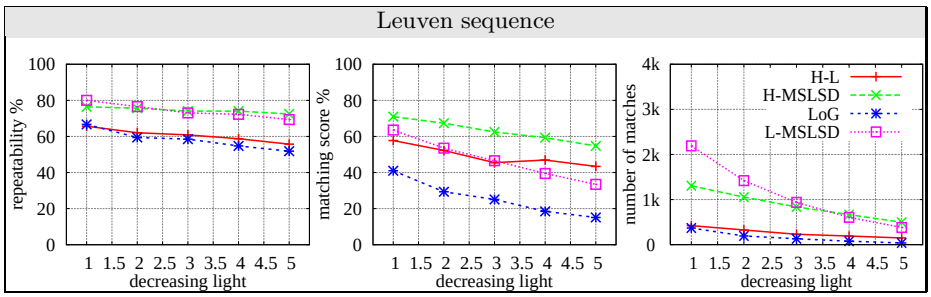

Fig. 7. Comparison of the detectors on the Leuven sequence (illumination changes).

the repeatability rate and matching score are significantly higher for our Harris- and Laplacian-based detectors than for the original H-L and LoG. This confirms that our scale selection is robust to lighting conditions as it is based on the SIFT descriptor which is invariant to affine illumination changes.

3.3 Overall performance

Mikolajczyk *et al.* [11] reported MSER (Maximally Stable Extremal Regions [14]) as the best affine-invariant detector on the three image sequences used here. Fig. 8 compares the matching score of our detectors to the performance of MSER on these sequences. Note that our results are directly comparable to the other

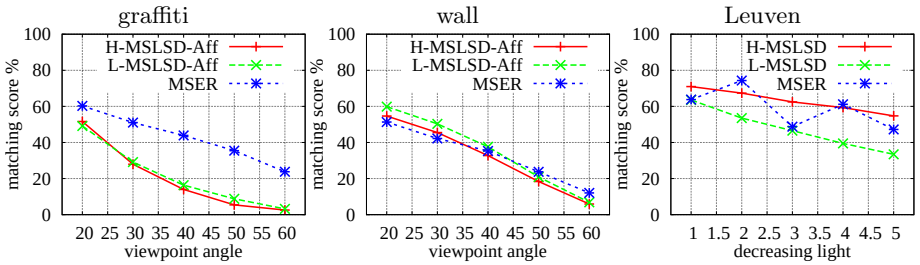

Fig. 8. Comparison of the matching scores obtained for our detectors, H-MSLSD-Aff and L-MSLSD-Aff, and MSER.

detectors reported in [11], as we use the same dataset and evaluation criteria. We can observe that L-MSLSD outperforms MSER on the wall sequence and that H-MSLSD performs better that MSER on the Leuven sequence. MSER gives better results than other detectors on the graffiti images. Note that due to the image structure of the graffiti scenes MSER selects significantly fewer keypoints than the other detectors.

4 Evaluation for image categorization

Category	H-L	H-MSLSD	LoG	L-MSLSD	Fergus *et al.* [1]	Opelt *et al.* [2]
Caltech databases						
Motorbikes	98.25	98.5	**98.75**	**98.75**	96.0	92.2
Airplanes	97.75	98.25	**99.0**	**99.0**	94.0	90.2
TUGraz1 databases						
Bicycles	92.0	**94.0**	90.0	92.0	*n.a.*	86.5
People	**86.0**	**86.0**	78.0	80.0	*n.a.*	80.8

Table 1. Comparison of object category classification results using our detectors (H-MSLSD and L-MSLSD) and their standard versions (H-L and LoG). Classification rates for four categories are reported at EER.

In this section we evaluate our new detectors for object and texture categorization. In both cases we perform image classification based on the bag-of-kepoints approach [18]. Images are represented as histograms of visual word occurrences, where the visual words are clusters of local descriptors. The histograms of the training images are used to train a linear SVM classifier. In the case of object categorization the output of the SVM determines the presence or absence of a category in a test image. For multi-class texture classification we use the 1-vs-1 strategy. Vocabularies are constructed by the K-Means algorithm. The number of clusters is fixed for each category, i.e., does not depend on the detector (400 for motorbikes and airplanes, 200 for bicycles, 100 for people, 1120

Database	H-L-R	H-MSLSD-R	LoG-R	L-MSLSD-R
Brodatz	$88.3_{\pm 0.6}$	$\mathbf{92.0}_{\pm \mathbf{0.5}}$	$90.5_{\pm 0.5}$	$\mathbf{95.8}_{\pm \mathbf{0.4}}$
KTH-TIPS	$83.9_{\pm 1.1}$	$\mathbf{88.4}_{\pm \mathbf{0.9}}$	$71.2_{\pm 1.5}$	$\mathbf{81.1}_{\pm \mathbf{1.2}}$

Table 2. Multi-class texture classification for two different datasets. The columns give the results for different detectors, here their rotation invariant versions.

for Brodatz, and 1000 for KTH-TIPS). In all experiments we compare H-L to H-MSLSD and LoG to L-MSLSD and our representation is always SIFT.

Evaluation for category classification. The experiments are performed for four different datasets. Motorbikes and airplanes of the CalTech dataset [1] contain 800 images of objects and 900 images of background. Half of the sets are used for training and the other half for testing. The split of the positive sets is exactly the same as [1]. The TUGRAZ-1 dataset [2] contains people, bicycles, and a background class. We use the same training and test sets for two-class classification as [2].

Table 1 reports the classification rate at the EER[1] for four databases and four different detectors. The last two columns give results from the literature. We can observe that in most cases our detectors give better results when compared to their standard versions. In the remaining cases the results are exactly the same. This demonstrates that the local description based on our detectors is more stable and representative of the data.

Evaluation for texture classification. Experiments are carried out on two different texture databases: Brodatz [19] and KTH-TIPS [20]. The Brodatz dataset consists of 112 different texture images, each of which is divided into 9 non-overlapping sub-images. The KTH-TIPS texture dataset contains 10 texture classes with 81 images per class. Images are captured at 9 scales, viewed under three different illumination directions and three different poses. Our training set contains 3 sub-images per class for Brodatz and 40 images per class for KTH-TIPS. Each experiment is repeated 400 times using different random splits and results are reported as the average accuracy on the folds with their standard deviation over the 400 runs. Table 2 compares the results of our detectors H-MSLSD-R and L-MSLSD-R to H-L-R and LoG-R. Note that we use the rotation invariant version here, as rotation invariance allows to group similar texture structures. We can observe that our scale selection technique, MSLSD, improves the results significantly in all cases.

Table 3 analyzes the influence of rotation invariance on the representation. Results for Harris-Laplace and LoG are in general better *without*, whereas results for our detectors are always better *with* rotation invariance. The poor performance of the existing detectors is due to an unstable estimation of the orientation leading to significant errors/noise in the descriptions. Note that the orientation of the patch is estimated after the region detection. In our MSLSD method rotation estimation is integrated into the scale selection criterion which

[1] Point on the ROC curves for which $p(TruePositives) = 1 - p(FalsePositives)$.

Brodatz		
Detector	no rot.inv.	rot.inv. (-R)
H-L	$89.2_{\pm 0.6}$ ⇐	$88.3_{\pm 0.6}$
H-MSLSD	$91.5_{\pm 0.6}$ ⇒	$92.0_{\pm 0.5}$
LoG	$90.1_{\pm 0.5}$ ⇒	$90.5_{\pm 0.5}$
L-MSLSD	$94.2_{\pm 0.5}$ ⇒	$95.8_{\pm 0.4}$

(a)

KTH-TIPS		
Detector	no rot.inv.	rot.inv. (-R)
H-L	$85.8_{\pm 1.1}$ ⇐	$83.9_{\pm 1.1}$
H-MSLSD	$88.1_{\pm 1.2}$ ⇒	$88.4_{\pm 0.9}$
LoG	$73.1_{\pm 1.5}$ ⇐	$71.2_{\pm 1.5}$
L-MSLSD	$80.9_{\pm 1.3}$ ⇒	$81.1_{\pm 1.2}$

(b)

Table 3. Classification accuracy with and without rotation invariance. Results for the Brodatz (a) and KTH-TIPS (b) datasets and different detectors.

implies that only regions with stable dominant gradients are selected, and it therefore improves the quality of the image representation.

5 Conclusion and future work

This paper introduced a new approach for selecting characteristic scales based on the stability of the local description. We experimentally evaluated this technique for the SIFT descriptor, i.e. Maximally Stable Local SIFT Description (MSLSD). We also demonstrated how a stable estimate of affine regions and orientation can be integrated in our method. Results for MSLSD versions of Harris and Laplacian points outperformed in many cases their corresponding state-of-the-art versions with respect to repeatability and matching. For object category classification MSLSD achieved better or similar results for four datasets. In the context of texture classification our approach always outperformed the standard versions of the detectors.

Future work includes the evaluation of our maximally stable local description approach with other keypoint detectors as well as other descriptors. Our scale selection could also be applied to a dense image representation, which would require an additional criterion for selecting discriminative regions.

Acknowledgments

This research was supported by the European project LAVA.

References

1. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR, Madison, Wisconsin, USA. Volume II. (2003) 264–271
2. Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: Weak hypotheses and boosting for generic object detection and recognition. In: ECCV, Prague, Czech Republic. Volume II. (2004) 71–84
3. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets. In: ECCV, Copenhagen, Denmark. Volume I. (2002) 414–431

4. Lazebnik, S., Schmid, C., Ponce, J.: Sparse texture representation using affine-invariant neighborhoods. In: CVPR, Madison, Wisconsin, USA. Volume 2. (2003) 319–324
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
6. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: CVPR, Madison, Wisconsin, USA. Volume 2. (2003) 257–263
7. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60**(1) (2004) 63–86
8. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference. (1988) 147–151
9. Lindeberg, T., Garding, J.: Shape-adapted smoothing in estimation of 3D depth cues from affine distortions of local 2D brightness structure. In: ECCV, Stockholm, Sweden. (1994) 389–400
10. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: ECCV, Prague, Czech Republic. Volume I. (2004) 228–241
11. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65**(1/2) (2005) 43–72
12. Tuytelaars, T., Van Gool, L.: Matching widely separated views based on affine invariant regions. IJCV **59**(1) (2004) 61–85
13. Jurie, F., Schmid, C.: Scale-invariant shape features for recognition of object categories. In: CVPR, Washington, DC, USA. Volume II. (2004) 90–96
14. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC, Cardiff, England. (2002) 384–393
15. Blostein, D., Ahuja, N.: A multi-scale region detector. Computer Vision, Graphics and Image Processing **45**(1) (1989) 22–41
16. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision **30**(2) (1998) 79–116
17. Baumberg, A.: Reliable feature matching across widely separated views. In: CVPR, Hilton Head Island, South Carolina, USA. Volume I. (2000) 774–781
18. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. (2004) 1–22
19. Brodatz, P.: Textures: A Photographic Album for Artists and Designers. Dover Publications, New York (1966)
20. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.O.: On the significance of real-world conditions for material classification. In: ECCV, Prague, Czech Republic. Volume IV. (2004) 253–266