# A comparison of local feature detectors and descriptors for visual object categorization by intra-class repeatability and matching

Jukka Lankinen, Ville Kangas, and Joni-Kristian Kamarainen
*Machine Vision and Pattern Recognition Laboratory* (`http://www2.it.lut.fi/mvpr`)
*Lappeenranta University of Technology (LUT), Kouvola Unit*

## Abstract

*Intuitive and easily interpretable performance measures, repeatability and matching performance, for local feature detectors and descriptors were introduced by Mikolajczyk et al. [10, 9]. They, however, measured performance in a wide baseline setting that does not correspond to the visual object categorisation problem which is a popular application of the detectors and descriptors. The limitation has been recognised and ad hoc evaluations proposed. To the authors' best knowledge, our work is the first which extends the original repeatability and matching performance measures to the case of object classes. Using the novel evaluation framework we test state-of-the-art detectors and descriptors with the popular Caltech-101 dataset and report the object category level (intra-class) repeatability and matching performances.*

## 1. Introduction

Local feature detectors and descriptors are popular in computer vision applications, where feature correspondences between two or more images are needed. The detectors and descriptors should tolerate illumination changes, zoom, blur and other typical distortions. This is the case in many applications, such as in wide-baseline matching [13], robot localisation [12] and image stitching for panoramic views [2]. In these, the correspondences are sought between different views of a same scene and the results in [10, 9] help to select the most suitable method. Another popular application is visual object categorisation, where objects in images should be automatically identified. In this case, the evaluations in [9] and [10] are not directly applicable, since the repeatability and matching were evaluated for multiple views of a same object (scene), not for multiple images of a visual object class.

Various methods have been proposed for detecting interest points/regions and to construct descriptors for them. Most of which are designed with a different application in mind. In [10] and [9] Mikolajczyk et al. evaluated and compared the most popular detectors and descriptors. The detectors were evaluated by their repeatability ratios and total number of correspondences for different view points of several views and with various imaging distortions. The descriptors were evaluated by their matching rates for the same scenes. In this work, we evaluate state-of-the-art detectors and descriptors in the visual object categorisation context and make the following important contributions:

- We extend the detector repeatability evaluation procedure in [10] for object categories. The intra-class number of correspondences and repeatability rates are reported as the performance numbers.
- We extend the descriptor matching evaluation in [9] for object categories. The intra-class match counts/rates are reported.
- We compare a set of the most popular detector and descriptor methods and their various implementations in our novel intra-class setting.

### 1.1 Restrictions and related work

We believe that the general evaluation principles in [9, 10] also hold in the visual object categorisation context: 1) *detectors which return the same object regions for category examples are good detectors* – detection repeatability; 2) *descriptors which match the same object regions between category examples are good descriptors* – match count/ratio. Success in the final task, categorisation, is important for the final application, and therefore Zhang et al. [14] compared different detectors and descriptors using a baseline bag-of-features (BoF) method. In their work, Mikolajczyk et al. [7] were more specific by measuring average precision of feature clusters to represent a single class, entropy of spatial location distribution produced by a single cluster (ideally compact) and complementarity of different detectors. The both evaluations are biased by the fixed approach: BoF. In this work, we adopt the original evaluation principles and thus obtain quantitative performance in the general and intuitive terms used in the original works, and not tied to any specific approach.

## 2 Comparison of Region Detectors

A good detector should detect local points or regions at the same relative locations, "object landmarks", on every class example. This criterion is different from Mikolajczyk et al. [10] in the sense that we evaluate detectors over different instances instead of different views. In our case, visual appearance variation is expected to be much larger.

### 2.1 Data

The experiments were conducted using the popular Caltech-101 dataset [3]. We report the results for the following ten categories which represent well the overall variation in the dataset: *watch*, *stop_sign*, *starfish*, *revolver*, *euphonium*, *dollar_bill*, *car_side*, *airplanes*, *Motorbikes* and *Faces_easy*. The provided foreground areas were used to mask out interested points detected on the backgrounds. Affine correspondences between the examples were established by manually annotating at least 5 object landmarks and by estimating the pairwise transformations with the direct linear transform [4] (see Fig. 1). For this experiment we used 25 random pairs of images from each category (tot. of 500 images).

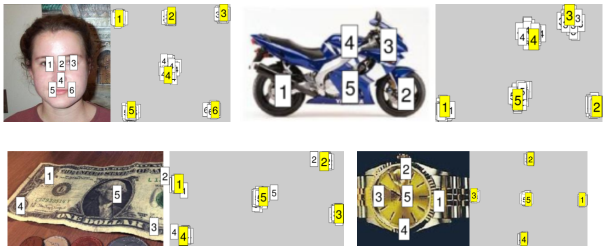


Figure 1: Object class examples and annotated landmarks. Also all object landmarks projected onto the first example (denoted by the yellow tags). The two standard deviations of the image diagonal normalised projection errors varied between 0.0158 and 0.0641.

### 2.2 Region detectors

Our comparison includes nine publicly available and popular detectors, which have performed best in the earlier studies. Hessian-affine detector [8] performed well in the comparison by Mikolajczyk et al. [10], and we included Mikolajczyk's original (*hessaff-alt*) and a more recent implementation (*hessaff*), and an alternative by Zhao [15] (*hesslap-vireo*). Our set of "fast detectors" consisted of Difference-of-Gaussian (DoG) by Lowe [5] (*sift*), Zhao's implementation of DoG (*dog-vireo*) and speeded-up robust features (SURF) by Bay et al. [1] (*surf*). In addition, Zhao's implementation of Laplacian-of-Gaussian (LoG) (*log-vireo*), Harris-Laplace (*harlap-vireo*) and Maximally Stable Extremal Regions (MSER) by Matas et al. [6] (*mser*) were included. Experiments were conducted using the available implementations with their default parameters.

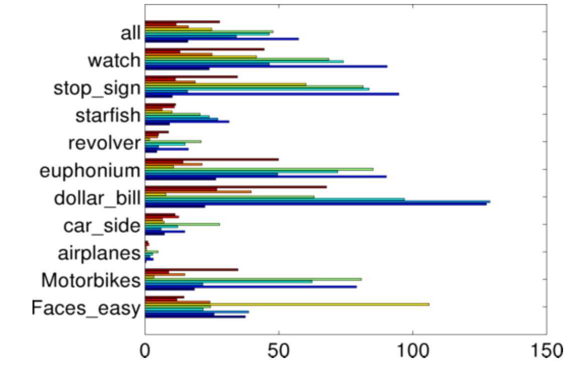
### 2.3 Performance evaluation

For the detector performance evaluation, we adopted the test protocol in [10]. Interest points are first extracted from images. The points with their centroid in the object area (Caltech-101 foreground) are selected for the evaluation. For each image pair, points from the first image are projected onto the second image. The projection is affine transformation estimated using the annotated landmarks. The landmarks projected on the first example of each category are demonstrated in Figs. 1 with the two standard deviations corresponding to the 95% error distributions. Interest regions are described by 2D ellipses and a sufficient overlap of fixed scale ellipses in the both images is accepted as a correct correspondence [10]. The number and rate of the correspondences for each detector is of interest. A detector performs well if the total number is large and reliable if also the ratio is high. We used the parameter settings from [10]: 40% overlap threshold and normalisation of the ellipses to the radius of 30 pixels.

### 2.4 Results

Results are gathered to Fig. 2. There are significant differences between the different categories. Dollar bill and stop sign are generally the easiest, as expected due to low variability in their visual appearance, while the airplanes, car side views and revolvers are the most difficult. The numbers of regions are by order of magnitude smaller than for the fixed scenes in [10], being tens of correspondences instead of hundreds of them.

The following three methods have good repeatability ratio: hesslap-vireo, dog-vireo and surf. Hesslap-vireo ($\approx$ hessaff), is very good as its repeatability rate is the third best (30%) and it provides more correspondences (57) on average. Dog-vireo ($\approx$ sift) has the best repeatability (33%), but its average number of correspondences (16) can be too low for large scale categorisation. Since the repeatability rates are very similar for the three best, the selection is based on the preferred number of correspondences. As a summary, the recent implementations: hesslap-vireo, hessaff and log-vireo perform best,



(a)

(b)

| Method | Avg # of corr. | rep. rate |
|---|---|---|
| *dog-vireo* | 16.0 | 33.7% |
| *hesslap-vireo* | 57.4 | 30.6% |
| *harlap-vireo* | 34.2 | 20.3% |
| *log-vireo* | 46.5 | 26.3% |
| *hesaff* | 47.8 | 25.3% |
| *hesaff-alt* | 25.0 | 23.4% |
| *sift* | 16.2 | 21.5% |
| *mser* | 11.7 | 13.8% |
| *surf* | 27.9 | 32.0% |

(c)

Figure 2: Detector evaluation: (a) average number of corresponding regions, (b) colour coding of the methods, and (c) overall results table (inc. repeatability rate).

≈ 50 corresponding regions on average with 25-30% repeatability rate.

# 3 Comparison of Region Descriptors

A good region descriptor for the categorisation problem should be discriminative to match only correct regions, but also tolerate small appearance variation between category examples.

## 3.1 Selected descriptors

Out of the many available descriptors we wanted to test the most frequently used and best performing. In the original comparison [9], SIFT detector by Lowe [5] and its extension, the gradient location and orientation histogram (GLOH) [9], obtained the best results. SIFT was selected for this comparison. A more recent descriptor SURF, besides of being very fast detector, is a robust descriptor and conceivably more tolerant to at least moderate amount of noise, than SIFT [1]. In the most works, the used descriptors are within these three. Moreover, we implemented one “traditional” descriptor, response vector of oriented filters, included in the original work [9] (steerable filters).

For the SIFT descriptor, we selected two implementations, the original by Lowe [5] (*sift*) and a more recent by Zhao [15] (*sift-vireo*). The orientation filter descriptor (*lin.filters*) was our own implementation. For SURF the implementation by its original authors [1] (*surf*) was used. Ideally, we should combine these descriptors with all the three best detectors, but the following best combinations were selected according to our preliminary tests: 1) *hesslap-vireo+sift-vireo* (≈ hesaff+sift), 2) *dog-vireo+sift-vireo* (≈ sift+sift), 3) *hesaff+sift*, 4) *surf+surf*, 5) *hesaff+lin.filters* and 6) *sift+sift*. It should be noted, that the available executable do not allow arbitrary combinations.

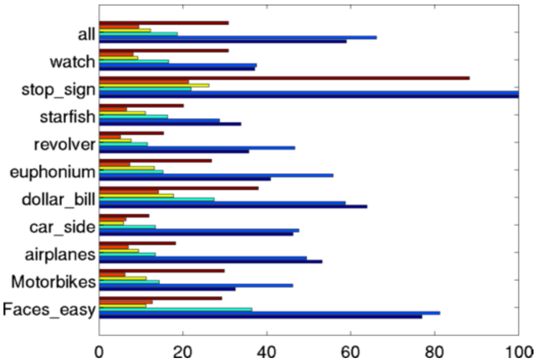
## 3.2 Performance evaluation

The main work flow is similar to [9]. At first, descriptors are computed for all detected regions (foreground only). Images are processed pair-wise and the best matches for the regions sought by computing one-to-all distances and selecting the closest match.

Our spatial verification stage differs from [9] by being less strict since the original rule provides only a few matches for the most pairs. In the original rule, the regions were described by ellipses, and for the spatial verification the ellipses were projected onto each other using the estimated affine transformation. If a sufficient overlap occurred for the ellipses, then the match was accepted. In our case, however, the categories have natural variation in their spatial structure. This natural variation cannot be exactly encoded into affine transformation and therefore the matches are not exact even for the ground truth landmarks as demonstrated in Fig. 1. The two standard deviations vary between 0.0158 (Faces_easy) and 0.0641 (euphonium). However, for ellipse overlap computation, even a small difference in the ellipse centroid may have an enormous effect to the overlap area [10]. We replaced the ellipse overlap rule with a distance threshold between the ellipse centroids. For resolution independence, the distances were normalised with the image diagonal and in our evaluation we discarded matches if the distance was greater than 0.05. This threshold covers the two standard deviations of the ground truth landmarks, i.e. 95% of the landmarks are within this distance.

## 3.3 Results

The data for this experiment were the same. The average and median number of matches are shown in Fig. 3. Our results verify the findings in the earlier works: the Hessian-Affine and SIFT detector-descriptor-pair leads to the largest number of matches. Overall, these results also seem to verify the important finding by Nowak et al. [11] that detector-descriptor combinations with a detector providing a larger number of correspondence candidates perform well. *hessaff* and *hesslap* based methods clearly outperform those using *sift* (*dog*) and *surf* (see also Fig. 2(c)). The weaker performance of hesslap-vireo+hesslap-sift can be explained by the fact that the vireo code does not do full affine normalisation (only one iteration) which seems to degrade matching with the SIFT descriptor.



(a)

- 🟦 hesaff+lin. filters
- 🟦 hesaff+sift
- 🟦 dog−vireo+sift−vireo
- 🟨 surf+surf
- 🟧 sift+sift
- 🟥 hesslap−vireo+sift−vireo

(b)

| Method | Avg # | Med # |
|---|---|---|
| *hesaff+lin.filt.* | 58.9 | 39.0 |
| *hessaff+sift* | 66.1 | 46.0 |
| *dog-vireo+sift-vireo* | 18.7 | 15.0 |
| *surf+surf* | 12.3 | 9.0 |
| *sift+sift* | 9.5 | 6.0 |
| *hesslap-vireo+sift-vireo* | 30.9 | 22.0 |

(c)

Figure 3: Descriptor evaluation: (a) average number of matches per class, (c) colour coding, and (c) overall results table.

## 4 Conclusions

In this work, the good and intuitive interest point detector and descriptor performance measures by Mikolajczyk et al. [10, 9], repeatability and number of matches, were extended to measure intra-class performance with visual object categories. The most popular state-of-the-art detectors and descriptors were compared using the Caltech-101 data set. This work was motivated by the fact that the original works studied a wide baseline setting which does not correspond to use of the detectors and descriptors in visual class detection and categorisation; tasks with clearly distinct requirements.

The detector experiment showed that SIFT and SURF are the most reliable in the terms of repeatability rate for object categories, but their marginal to the Hessian-affine is not significant and the Hessian-affine provides more interest points (48 vs. 16 for the SIFT and 28 for the SURF) as a complementary feature.

The descriptor experiment proved that descriptors paired with the Hessian-affine detector perform best for matching similar regions over multiple examples of a same object class. Hessian-affine detector (the latest implementation) and SIFT descriptor provided clearly the best results by average number of matches 66 and median 46.

Our results indicate that the Hessian-affine and SIFT form the best combination for object classification methods using local feature detectors and descriptors.

## References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[2] M. Brown and D. Lowe. Recognising panoramas. In *ICCV*, pages 1218–1227, 2003.

[3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE PAMI*, 28(4):594, 2006.

[4] R. Hartley and A. Zisserman. *Multiple View Geometry in computer vision*. Cambridge press, 2003.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. In *Int J Comput Vis*, volume 60, pages 91–110, 2004.

[6] J. Matas., O. Chum, M. Urban, and T. Padja. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[7] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *CVPR*, 2005.

[8] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, 2002.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 27(10):1615–1630, 2005.

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

[11] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[12] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *Int'l Conf. of Intelligent Robots and Systems*, pages 226–231, 2002.

[13] T. Tuytelaars and L. van Gool. Matching widely separated views based on affine invariant regions. *Int J Comput Vis*, 1(59), 2004.

[14] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int J Comput Vis*, 73(2), 2006.

[15] W. Zhao. LIP-VIREO local interest point extraction toolkit. http://vireo.cs.cityu.edu.hk.