Elsevier Editorial System(tm) for Neurocomputing
Manuscript Draft

Manuscript Number:

Title: A Comparison of Feature Detectors and Descriptors for Object Class Matching

Article Type: SI: Robust local descriptors

Keywords: local descriptor; local detector; interest point; SIFT; SURF;
BRIEF; BRISK; ORB; FREAK

Corresponding Author: Dr. Joni-Kristian Kämäräinen, PhD

Corresponding Author's Institution: Tampere University of Technology

First Author: Antti Hietanen, M.Sc.

Order of Authors: Antti Hietanen, M.Sc.; Jukka Lankinen, PhD; Anders G Buch, PhD; Joni-Kristian Kämäräinen, PhD; Norbert Kruger, PhD

Abstract: Solid protocols to benchmark local feature detectors and descriptors were introduced by Mikolajczyk et al. 2005. The detectors and descriptors are popular tools in object class matching, but the wide baseline setting in the benchmarks
does not correspond to class-level matching where appearance variation can be large. We extend the benchmarks to the class matching setting and evaluate state-of-the-art detectors and descriptors with Caltech and ImageNet classes.
Our experiments provide the following interesting findings with regard to object class matching: 1) the original SIFT is still the best descriptor; 2) dense sampling outperforms interest point detectors with a clear margin; 3) detectors
perform moderately well, but descriptors' performance collapse; 4) using multiple, even a few, best matches instead of the single best has significant effect on
the performance; 5) object pose variation degrades dense sampling performance while the best detector (Hessian-affine) is unaffected. The performance of the best detector-descriptor pair is verified in the application of unsupervised visual class alignment where state-of-the-art results are achieved. The findings help to improve the existing detectors and descriptors for which the framework provides an automatic validation tool.

A Comparison of Feature Detectors and Descriptors for Object Class Matching

Antti Hietanen, Jukka Lankinen, Joni-Kristian Kämäräinen[1]

Department of Signal Processing, Tampere University of Technology

Anders Glent Buch, Norbert Krüger

Maersk Mc-Kinney Moller Institute, University of Southern Denmark

Abstract

Solid protocols to benchmark local feature detectors and descriptors were introduced by Mikolajczyk et al. [1, 2]. The detectors and descriptors are popular tools in object class matching, but the wide baseline setting in the benchmarks does not correspond to class-level matching where appearance variation can be large. We extend the benchmarks to the class matching setting and evaluate state-of-the-art detectors and descriptors with Caltech and ImageNet classes. Our experiments provide important findings with regard to object class matching: 1) the original SIFT is still the best descriptor; 2) dense sampling outperforms interest point detectors with a clear margin; 3) detectors perform moderately well, but descriptors' performance collapse; 4) using multiple, even a few, best matches instead of the single best has significant effect on the performance; 5) object pose variation degrades dense sampling performance while the best detector (Hessian-affine) is unaffected. The performance of the best detector-descriptor pair is verified in the application of unsupervised visual class alignment where state-of-the-art results are achieved. The findings help to improve the existing detectors and descriptors for which the framework provides an automatic validation tool.

Keywords: local descriptor, local detector, interest point, SIFT, SURF,

[1]joni.kamarainen@tut.fi; +358 50 300 1851; P.O.Box 553, FI-33101 Tampere, Finland

1. Introduction

Image feature detectors and descriptors are the tools in computer vision problems where point or region correspondences between images are needed. Ideally, they should tolerate pose variation, illumination changes, motion blur and other typical scene changes and distortions. That is the case, for example, in wide baseline matching [3], robot localization [4] and panorama image stitching [5]. In these cases, the feature correspondences are needed to match several views of same scenes and the detector and descriptor evaluations by Mikolajczyk and Schmid 2005 [1] and Mikolajczyk et al. 2005 [2] help to find the most suitable detector-descriptor pair. A distinct application of feature-based matching is visual object classification and detection, where instances of object classes must be identified and localized in input images. In that case, the visual appearance variation can be very large as compared to fixed scenes, and thus, the original evaluations are not directly applicable.

Figure 1: Numbers of descriptor matches between two random class examples.

Various methods have been proposed for detecting interest points/regions and to construct descriptors from them, most of which are designed with a different application in mind. Recently, fast detectors and descriptors have been proposed: SURF [6], FREAK [7], ORB [8], BRISK [9], BRIEF [10]. In [1] detectors were evaluated by their repeatability ratios and total number of correspondences over several views of scenes and with various imaging distortion types. In [2] descriptors were evaluated by their matching rates for the same views. Comparisons on object classification were reported in [11] and [12], but they were tied to a single approach, visual Bag-of-Words (BoW). Our main contributions are:

- We introduce intuitive detector and descriptor evaluation frameworks by extending the detector and descriptor benchmarks in [1, 2] to intra-class repeatibility and matching.
- We evaluate the recent and popular detectors and descriptors and their various implementations with the proposed framework.
- We investigate the effect of using multiple best matches ($K = 1, 2, \ldots$) and introduce an alternative performance measure: *match coverage*.

From the experimental results on Caltech and ImageNet classes we arrive at the following important findings:

- Dense SIFT features are the best.
- Detectors generally perform well, but the ability of descriptors to match regions over visual class examples is poor (Fig. 1).
- Using multiple—even a few—best matches instead of the single best provides significant improvement.
- Dense grid sampling outperforms interest point detectors with a clear margin, but
- object pose variation can drastically affect dense sampling while the best detector (Hessian-affine) is unaffected.
- The original SIFT is still the best descriptor.

Source code for the evaluation framework will be published in the Web[2]. In addition, we verify our findings with the application of unsupervised object class alignment where the best detector-descriptor pair improves the state-of-the-art.

1.1. Related work

We believe that the general evaluation principles in [1, 2] also hold in the context of visual object classes: 1) *detectors which return the same object regions for class examples are good detectors* – detection repeatability; 2) *descriptors which match the same object regions between class examples are good descriptors* – match count/ratio. We refer to these repeating and matching regions as "category-specific landmarks". A qualitative measure to visualize descriptors ("HOGgles") was recently proposed by Vondrick et al. [13], but its main use is in visualization. More quantitative evaluations were reported by Zhang et al. [11] and Mikolajczyk et al. [12], but these were tied to a single methodology, the visual Bag-of-Words (BoW) [14, 15]. In this work, we show that the original evaluation principles can be adopted to obtain similar quantitative performance measures in general, comparable and intuitive forms to the original works of Mikolajczyk et al., and not tied to any specific approach.

2. Comparing Detectors

A good feature detector should detect local points or regions at the same locations of class examples to make it possible to match corresponding "parts". This criterion differs from [1], where detectors were evaluated over views of same scenes corresponding to specific object matching. In part-based object classification (e.g., [16]), the descriptors (parts) should match despite substantial variance in their visual appearance.

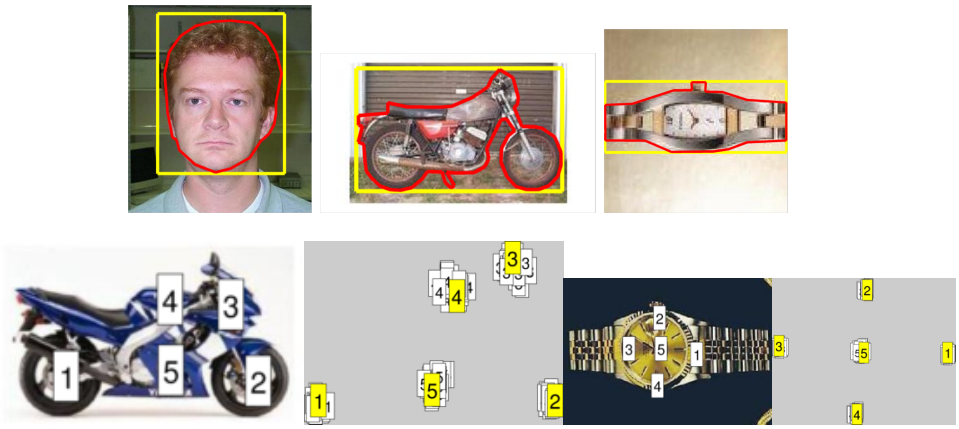


Figure 2: Top: example images with the provided ground truth (bounding boxes and foreground regions). Bottom: landmark examples and multiple landmarks projected onto a single image (the yellow tags).

2.1. Data

The experiments were conducted with the Caltech-101 [17] images. Caltech-101 is preferred as the baseline since objects' poses are roughly fixed that allows us to measure the effect of appearance variation without geometric pose noise. In the additional experiments we verify our results with randomly rotated versions of the Caltech images and the recent ImageNet database [18]. The foreground masks were used to remove features detected in the background (Fig. 2). Affine correspondence between category examples were established by manually annotating 5-12 landmarks per category and estimating the pair-wise image transformations using the direct linear transform [19] and linear interpolation. 25 random pairs from each class were repeatedly picked.

2.2. Feature detectors

The detectors for the experiments were selected among the best performing from our preliminary study [20] and the recently proposed detectors: BRIEF [10], BRISK [9], ORB [8] and FREAK [7]. The preliminary detectors were

[2] `https://bitbucket.org/kamarain/descriptor_vocbenchmark`

1. Two implementations of the difference of Gaussian: *sift* and *dog-vireo*
2. Harris-Laplace: *harlap-vireo*
3. Laplacian of Gaussian (log): *log-vireo*
4. Three implementations of the Hessian-affine: *hessaff*, *hessaff-alt* and *hesslap-vireo*
5. Speeded-up robust features: *surf*
6. Maximally stable extremal regions: *mser*

The detectors are publicly available: **-vireo* implementations in Zhao's Lip-vireo toolkit (`http://code.google.com/p/lip-vireo`), *hessaff* and *hessaff-alt* (by Mikolajczyk) at `http://featurespace.org`, *surf* at the authors' [6] web site and *mser* and *sift* in the VLFeat toolbox (`http://vlfeat.org`). The best average repeatability was 33.7% for *dog-vireo* and the best number of corresponding regions 57.4 for *hesslap-vireo*. The best three detectors based on the both repeatability and number of regions were *hesslap-vireo* (30.6%, 57.4), *hessaff* (25.3%, 47.8) and log-vireo (26.3%, 46.5). We report results for the best: the hessaff detector.

The best result from the recent detectors was obtained with the ORB OpenCV implementation (`http://opencv.org`) which is included (*orb*). Moreover, dense sampling has replaced detectors in the top methods (Pascal VOC 2011 [21]) and we added the dense SIFT in VLFeat (`http://vlfeat.org`) to our evaluation (*dense*).

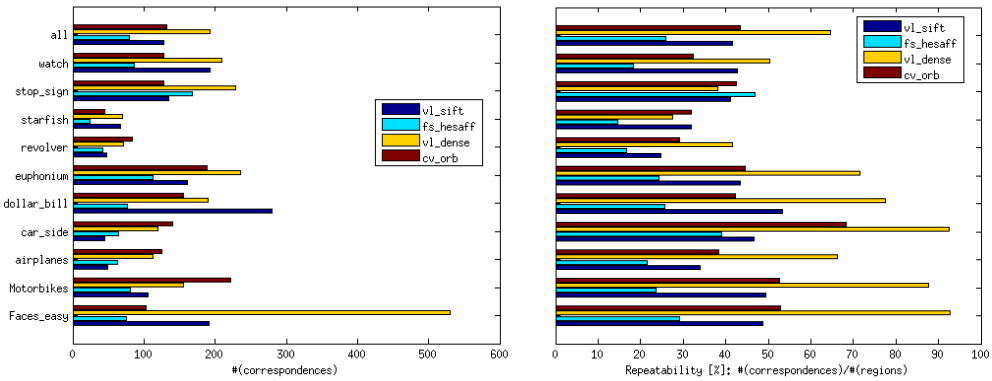
2.3. Performance measures and evaluation

For the detector performance evaluation, we adopted the procedure in [1] with the exception that interest points detected outside the object area (Fig. 2) are removed. For each image pair, points from the first image are projected onto the second image by the affine transformation estimated using the annotated landmarks. The interest points (regions) are described by 2D ellipses and when a transformed ellipse overlaps with an ellipse in the second image a correct correspondence is recorded. The number and rate of correspondences for each detector is of interest. A detector performs well if the total number is large and

has high precision if the ratio of correct matches is high. We used the parameter settings from [1]: 60% overlap threshold and normalization of the ellipses to the radius of 30 pixels. The normalization is required since the overlap area depends on the size of the ellipses.

The reported performance numbers are the average number of correspences between image pairs and the repeatability rate, i.e. the number of correspondences divided by the total number of points.

2.4. Results



(a) (b)

Detector	*Avg # of corr.*	*Avg. rep. rate*
vl_sift	127.5	41.6%
fs_hessaff	79.3	26.0%
cv_orb	132.0	43.5%
vl_dense	192.3	64.6%

(c)

Figure 3: Detector evaluation in object class matching. Meta-parameters were set to return on average 300 regions. (a) average number of corresponding regions, (b) repeatability rates, and (c) the overall results table.

It is noteworthy that this experiment differs from our preliminary work in the sense that instead of using the default parameters for each detector we

adjusted their meta-parameters to return on average 300 regions for each image (see Sec. 2.5 for further analysis). The results of the detector experiment are shown in Fig. 3. With the adjusted meta-parameters the difference between the detectors is less significant than in our preliminary work [20] and the previous winner, Hessian-affine, is now the weakest. The dense sampling is clearly better than others, but otherwise the ORB detector seems attempting due to its speed. It is also noteworthy that without the parameter adjustment the results of the original SIFT detector would be by order of magnitude worse. Some less favorable properties of dense sampling are discussed in Sec. 4.4.

2.5. Detecting more regions

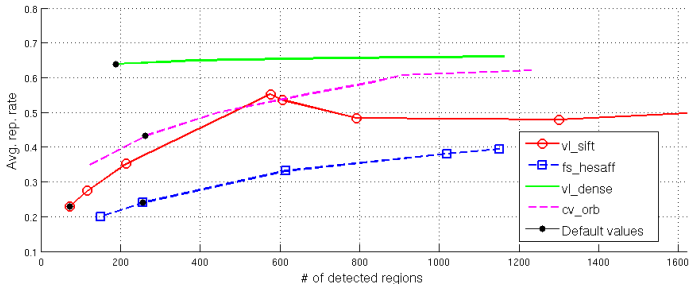


Figure 4: Detector repeatability as the function of the number of detected regions adjusted by the meta-parameters (defaults marked by black dots).

In the previous example, we adjusted detector meta-parameters to return on average 300 regions for each image. That made detectors produce very similar results while using the default parameters in our previous work lead to completely different interpretation. It is interesting to study whether we can exploit meta-parameters further to increase the number of corresponding regions. For ORB we adjusted the edge threshold, for Hessian-affine the feature density and the Hessian threshold, for SIFT the number of levels per octave, and for the dense the grid step size. We computed the detector repeatability rates as the functions of the number of detected regions (see Figure 4). As expected the

meta-parameters have almost no effect to the dense detection while Hessian-affine, ORB and especially SIFT clearly improve as the number of regions increase (SIFT regions saturate to the same locations approx. at 600 detected regions). For the most difficult classes in Fig. 3 (starfish and revolver) more regions is beneficial opening a novel research direction whether the detector parameters should be optimized for class detection?

3. Comparing Descriptors

A good region descriptor for object matching should be discriminative to match only correct regions, and also tolerate small appearance variation between the examples. The descriptor performances were obtained in the original work [2] by computing statistics of the correct and false matches. Between different class examples, descriptor matches are expected to be weaker due to increased appearance variation. For example, scooters and road bikes are both in the Caltech-101 motorbikes category, but their pair-wise similarity is much weaker than between two scooters or two road bikes.

3.1. Available descriptors

This experiment is conducted using detector-descriptor pairs. Our preliminary set of descriptors was:

1. Hessian-affine and SIFT
2. Hessian-affine and steerable filters
3. Vireo implementation of Hessian-affine and SIFT
4. Original SIFT detector and SIFT descriptor
5. Alternative (Vireo) implementation of SIFT and SIFT
6. SURF and SURF

With their default parameters the combinations 1) and 2) utilizing Mikolajczyk's implementation of Hessian-affine detector were clearly superior to other methods [20], but here we adjust the meta-parameters to return the same average number of regions (300).

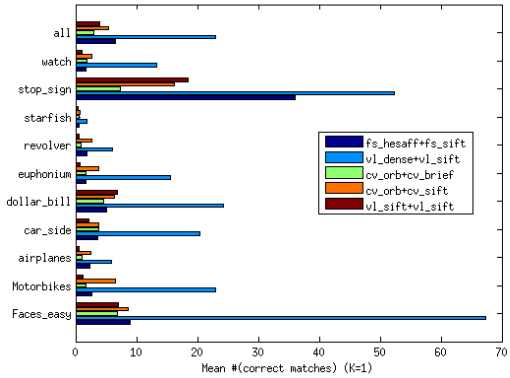
To these experiments, we also include the best fast detector-descriptor pair: ORB and BRIEF. The following combinations will be reported: *vl_sift+vl_sift* (FeatureSpace implementation), *fs_hessaff+fs_sift* (FeatureSpace implementation), *cv_orb+cv_brief* (OpenCV implementation), *cv_orb+cv_sift* (OpenCV, to compare SIFT and BRIEF), *vl_dense+vl_sift* (VLFeat implementation). We also tested the RootSIFT descriptor from [22] that achieved better performance in their experiments, but in our case it provided insignificant difference to the original SIFT (mean: $3.9 \rightarrow 4.2$, median: $1 \rightarrow 1$).

3.2. Performance measures and evaluation

In our preliminary work [20] we used a simplified version of the Mikolajczyk's descriptor performance measure: the ellipse overlap was replaced by normalized centroid distance of the matching regions. However, the results by the simplified rule turned out to be too optimistic and in this work we adopt the original measure. The rule is the same as with the detectors, if the best matching regions have sufficient overlap the match is counted correct. Descriptors are computed for all detected regions (foreground only). Images are processed pair-wise and the best match for each region is selected from the full distance matrix. It is worth noting that the rule proposed in [23] for discarding "bad regions" (ratio between the first and the second best is less than 1.5) is not used since it results complete failure. We used the ellipse overlap threshold 50% from [2], but also more strict thresholds were tested. Our performance numbers are the average number of matches and median number of matches. In the detector evaluation the mean and median numbers were almost the same, but here we report the both since for the descriptors there is significant discrepancies between the mean and median numbers.

3.3. Results

The average and median number of matches for the descriptor evaluation are shown in Fig. 5. For many classes, the mean and median numbers are very low and dense grid sampling is superior for all classes, achieving the average



(a)

Detector+descriptor	*Avg #*	*Med #*	*Avg # (60%)*	*(70%)*
vl_sift+vl_sift	3.9	1	2.8	1.6
fs_hessaff+fs_sift	6.5	2	5.9	4.9
vl_dense+vl_sift	23.0	10	22.3	20.2
cv_orb+cv_brief	3.0	1	2.9	2.7
cv_orb+cv_sift	5.4	2	4.8	4.1

(b)

Figure 5: Descriptor evaluation: (a) average number of matches per class, (b) overall results table. The default overlap threshold is 50% [2], 60% and 70% results demonstrate the effect of the more strict overlaps.

of 23.0 and median of 10.0 matches. The more strict overlaps, 60% and 70%, provide almost the same numbers verifying that the matched regions do match well also spatially.

The best results were obtained for the stop signs, dollar bills and faces, but the overall performance is poor. The best discriminative methods could still learn to detect these categories, but it is difficult to imagine naturally emerging "common codes" for other classes except the three. It is surprising that the best detectors, Hessian-affine and dense sampling, were able to provide 79 and 192, repeatable regions on average, but only roughly 10% of these match in

the descriptor space. Despite the fact that the SIFT detector performed well in the detector experiment, its regions do not match well in this experiment. The main conclusion of these results is that all descriptors perform poorly in matching regions between class examples.

3.4. The more the merrier?

Similar to Sec. 2.5 we study how the average number of matches behaves as the function of the number of extracted regions. This is justified as some works claim that "the more the merrier" [24]. The result graph is shown in Figure 6. The results show that adding more regions by adjusting the detector meta-parameters provides only minor improvement to the average number of matches. Clearly, the "best regions" are provided first and dense sampling performs much better indicating that what is "interesting" for the detectors is not necessarily a good object part.

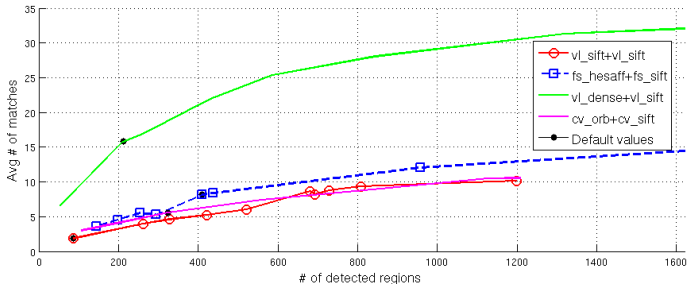


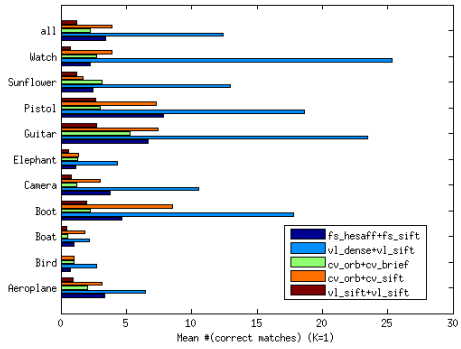
Figure 6: Descriptors' matches as functions of the number of detected regions controlled by the meta-parameters (default values denoted by black dots).

4. Advanced analysis

In this section, we address the open questions raised during the detector and descriptors comparisons in Section 2 and 3. The important questions are: why only a few matches are found between different class examples and what can be done to improve that? Why dense sampling outperforms all interest point

detectors and does it have any drawbacks? Do our results generalize to other datasets?

4.1. ImageNet classes



(a)

Detector+descriptor	*Avg #*	*Med #*	*Avg # (60%)*	*(70%)*
vl_sift+vl_sift	1.2	0	0.7	0.3
fs_hessaff+fs_sift	3.4	2	2.8	1.9
vl_dense+vl_sift	12.4	7	11.6	10.2
cv_orb+cv_brief	2.2	1	1.9	1.5
cv_orb+cv_sift	3.9	2	3.3	2.5

(b)

Figure 7: Descriptor evaluation with the ImageNet classes to verify results in Fig. 5.

To validate our results, we selected 10 different categories from the the state-of-the-art object detection database: ImageNet [18]. The images were scaled to the same size as the Caltech-101 images and the foreground areas were annotated. The results for the ImageNet classes are in Figure 7. The average number of matches is roughly half of the number of matches with Caltech-101 images which can be explained by the fact that the dataset is more challenging due to 3D view point changes. However, the ranking of the methods is almost the

same: dense sampling and SIFT is the best and SIFT detector and descriptor pair is the worst. The results validate our findings with Caltech-101.

4.2. Beyond the single best match

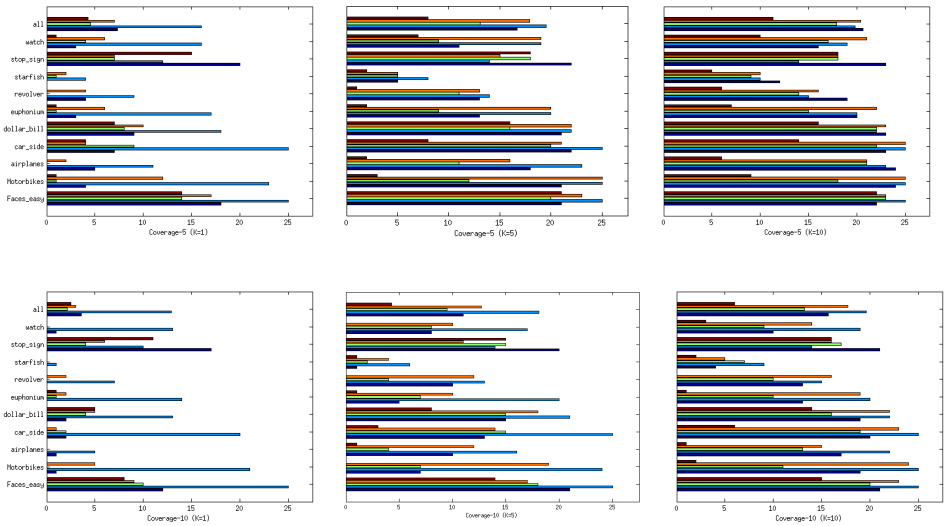


Figure 8: Number of image pairs for which at least $N = 5, 10$ (top,bottom) descriptor matches were found (*Coverage-N*). The best $K = 1, 5, 10$ (left-to-right) matches were counted.

In object matching, assigning each descriptor to several best matches, "soft assignment" [25, 26, 27], provides improvement and we want to experimentally verify this finding using our framework. To measure the effect of multiple assignments, we establish a new performance measure: *coverage*. Coverage corresponds to the number of image pairs for which at least N matches have been found (*coverage-N*) and this measure is more meaningful than the average number of matches since there was strong discrepancies between the average and median numbers. We tested the multiple assignment procedure by accumulating matches over $n = 1, 2, \ldots, K$ best matches. The corresponding coverage for $K = 1, 5, 10$ are shown in Fig. 8. Obviously, more image pairs contain at least $N = 5$ than $N = 10$ matches. With $K = 1$ (only the best match) the

best method, VLFeat dense SIFT, finds at least $N = 5$ matches in 16 out of 25 image pairs and 13 for $N = 10$. When the number of best matches is increased to $K = 5$, the same numbers are 19 and 18, respectively, showing clear improvement. Beyond $K = 5$ the positive effect diminishes and also the difference between the methods is less significant.

4.3. Different implementations of the dense SIFT

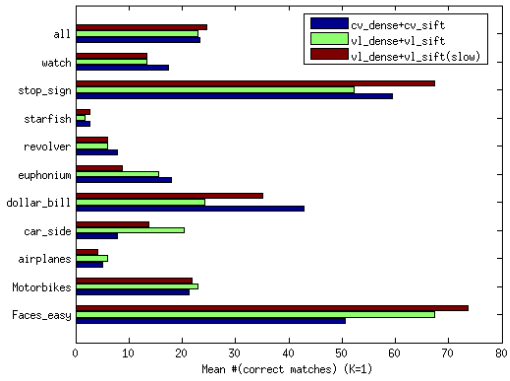


Figure 9: OpenCV dense SIFT vs. VLFeat dense SIFT (fast and slow) comparison.

During the course of work, we noticed that different implementations of the same method provided slightly different results. Since there are two popular implementations of dense sampling with the SIFT descriptor, OpenCV and VLFeat (two options: slow and fast), we compared them. The results corresponding to the previous experiments in Sec 3 are shown in Fig. 9. There are slight differences in classes due to implementation differences, but the overall performances are almost equal.

4.4. Challenging dense sampling: r-Caltech-101

In dense sampling the main concern is its robustness to changes in scale and, in particular, orientation, since these are not estimated similar to interest point detection methods. In this experiment, we replicated the previous experiments with the two dense sampling implementations and the best interest point

Figure 10: The r-Caltech-101 versions of the original Caltech-101 images in Fig. 2 (original bounding box shown by green).

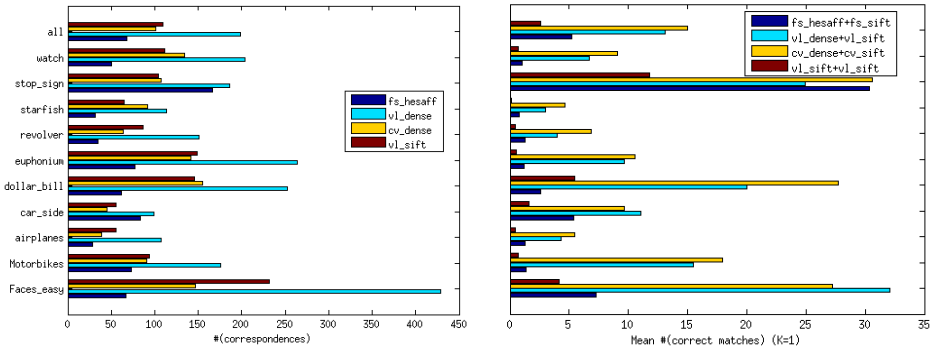


Figure 11: R-Caltech-101: detector (left) and descriptor (right). The detector results are almost equivalent to Fig. 3. In the descriptor benchmark (cf. with Fig. 5) the Hessian-affine performs better (mean: 3.4 $\rightarrow$ 5.2) while both dense implementations, VLFeat (23.0 $\rightarrow$ 13.1) and OpenCV (23.3 $\rightarrow$ 15.0) are severely affected.

detection method using the randomized version of the Caltech-101 data set, r-Caltech-101 [28]. R-Caltech-101 contains the same objects (foreground), but with varying random Google backgrounds and the objects have been translated, rotated and scaled randomly (Fig. 10).

The detector and descriptor results of this experiment are shown in Fig. 11 Now it is clear that artificial rotations affect the dense descriptors while Hessian-affine is unaffected (actually improves). It is noteworthy that the generated pose changes in r-Caltech-101 are rather small ($[-20°, +20°]$) and the performance drop could be more dramatic with larger variation. An intriguing research

direction is detection of scaling and rotation invariant dense interest points.

5. Application: Image alignment

Table 1: Unsupervised image alignment accuracy with the feature-based congealing [29] for the original setting (Hessian-affine+SIFT) and the best in our evaluation: dense SIFT.

						
Orig. [29]	78%	53%	76%	27%	24%	2%
Orig. optim.	88%	86%	78%	35%	24%	4%
Dense orig.	96%	71%	86%	86%	20%	65%
Dense optim.	**98%**	**92%**	**90%**	**92%**	**53%**	**76%**

To verify our findings in a real application where region detectors and descriptors are core tools we selected the unsupervised feature-based object class image alignment method [29] for which state-of-the-art alignment accuracy is reported. The method takes as inputs an image ensemble and a single image selected as a "seed". Matches between the seed and other images are computed and spatially matching seed descriptors accumulated over the process. The best seed descriptors are selected and all images are aligned using them. The process is simple, but depends on the success of the detector-descriptor pair. The original method uses the Hessian-affine detector and the SIFT descriptor with their default settings. The method's own meta-parameters are the normalized spatial match distance $\tau = 0.05$, the maximum number of seed landmarks $L = 20$, and the number of best descriptor matches $K = 10$.

The results are shown in Table 1 for the original detector-descriptor pair and for the vl_dense+vl_sift pair that performed best in our previous experiments. We used the same Caltech-101 classes from the previous experiments. During the experiments we found that the original parameter settings are sub-optimal and by cross-validation optimized them (Hessian-affine: $\tau = 0.02$, $L = 20$,

$K = 2$; dense: $\tau = 0.04$, $L = 80$, $K = 10$). The performance number is the proportion of correctly aligned images measured by the normalized average distance of the annotated landmarks after alignment (0.10 corresponds to 10% of the distance between the two furthest landmarks - "object size"). In the both original and optimized settings the dense SIFT is clearly superior to the Hessian-affine and provides much better alignment performance even for the classes for which the original method performs poorly (airplanes and revolvers) or fails (watches). See Figure 12 for alignment examples.

Figure 12: Average images without (top) and with unsupervised alignment (bottom).

6. Discussion

Interest points and regions have been the low-level features in visual class detection and classification for a decade [14]. Recently, supervised low-level features, such as convolution filters in deep neural networks [30], have gained momentum, but we believe that the unsupervised detector-descriptor approach can be developed further by identifying and improving the bottlenecks. In this work, we took a step to this direction by introducing an evaluation framework of part detectors and descriptors which provides intuitive and comparable results in the quantitative manner of the original works [1, 2].

With the proposed framework we identified the following important findings: 1) The original SIFT is the best descriptor (including the recent fast descriptors); 2) Dense sampling outperforms interest point detectors with a clear margin; 3) Detectors generally perform well, but descriptors' ability to match parts over visual class examples collapse; 4) Using multiple, even a few, best matches instead of the single best match provides significant performance boost; 5) Object pose variation severely affects dense sampling while the best detector (Hessian-affine) is almost unaffected.

The findings advocate new research on i) optimization of the detector meta-parameters per visual class, ii) specialized descriptors for visual class parts and regions, iii) dense scaling and rotation invariant interest points, and iv) alternative matching methods for multiple best matches. Some results already exist. For example, BoW codebook descriptors can be enhanced by merging descriptors based on co-location and co-activation clustering [31] or by learning [32], dense interest points have been proposed [33], and soft-assignment has been shown to improve BoW codebook matching [25]. Moreover, the success of the standard SIFT in our experiments justifies further development of more effective visual class descriptors, not only more efficient descriptors. Investigating these potential research directions benefits from our evaluation framework that can be used for automatic validation and optimization.

References

[1] 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.

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

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

[4] S. Se, D. Lowe, J. Little, Global localization using distinctive visual features, in: Int'l Conf. of Intelligent Robots and Systems, 2002, pp. 226–231.

[5] M. Brown, D. Lowe, Recognising panoramas, in: ICCV, 2003, pp. 1218–1227.

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

[7] A. Alahi, R. Ortiz, P. Vandergheynst, FREAK: Fast retina keypoint, in: CVPR, 2012.

[8] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: ICCV, 2011.

[9] S. Leutenegger, M. Chli, R. Siegwart, BRISK: Binary robust invariant scalable keypoints, in: ICCV, 2011.

[10] M. Calonder, V. Lepetit, C. Strecha, P. Fua, BRIEF: Binary robust independent elementary features, in: ECCV, 2010.

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

[12] K. Mikolajczyk, B. Leibe, B. Schiele, Local features for object class recognition, in: CVPR, 2005.

[13] C. Vondrick, A. Khosla, T. Malisiewicz, A. Torralba, HOGgles: Visualizing object detection features, in: ICCV, 2013.

[14] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV, 2003.

[15] G. Csurka, C. Dance, J. Willamowski, L. Fan, C. Bray, Visual categorization with bags of keypoints, in: ECCV Workshop on Statistical Learning in Computer Vision, 2004.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE PAMI 32 (9).

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

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR, 2009.

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

[20] J. Lankinen, V. Kangas, J.-K. Kamarainen, A comparison of local feature detectors and descriptors for visual object categorization by intra-class repeatability and matching, in: 21th Int. Conf. on Pattern Recognition (ICPR2012), 2012.

[21] M. Everingham, L. V. Gool, C. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results, http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html (2011).

[22] R. Arandjelovic, A. Zissermann, Three things everyone should know to improve object retrieval, in: CVPR, 2012.

[23] D. Lowe, Distinctive image features from scale-invariant keypoints, in: Int J Comput Vis, Vol. 60, 2004, pp. 91–110.

[24] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: ECCV, 2006.

[25] A. Agarwal, B. Triggs, Multilevel image coding with hyperfeatures, Int J Comput Vis 78 (1).

[26] T. Tuytelaars, C. Schmid, Vector quantizing feature space with a regular lattice, in: ICCV, 2007.

[27] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: BMVC, 2011.

[28] T. Kinnunen, J.-K. Kamarainen, L. Lensu, J. Lankinen, H. Kälviäinen, Making visual object categorization more challenging: Randomized Caltech-101 data set, in: 20th Int. Conf. on Pattern Recognition (ICPR2010), 2010.

[29] J. Lankinen, J.-K. Kamarainen, Local feature based unsupervised alignment of object class images, in: BMVC, 2011.

[30] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, 2012.

[31] B. Leibe, A. Ettlin, B. Schiele, Learning semantic object parts for object categorization, Image and Vision Computing 26 (2008) 15–26.

[32] K. Simonyan, A. Vedaldi, A. Zisserman, Learning local feature descriptors using convex optimisation, IEEE PAMI 36 (8).

[33] T.Tuytelaars, Dense interest points, in: CVPR, 2010.

***Biography of the author(s)**

Click here to download Biography of the author(s): biographies.txt

Prof. Joni-Kristian Kamarainen
Dr. Kamarainen received MSc (Eng.) and DSc (Tech.) degrees, both in information processing (Comp. Sc.), from Lappeenranta University of Technology in 1999 and 2003, respectively. He is a founding member and vice director of the Machine Vision and Pattern Recognition Laboratory, Lappeenranta University of Technology where he was appointed Professor of Information Society Technologies in 2008. His research interests include computer vision, image analysis and pattern recognition. He is a chairman of Pattern Recognition Society of Finland, and member of International Association for Pattern Recognition (IAPR).

Prof. Norbert Kruger
Norbert Kruger received the MSc degree from the Ruhr-Universitat Bochum, Germany, and the PhD degree from the University of Bielefeld. He is a professor at the Maersk McKinney Moller Institute, University of Southern Denmark. He leads the Cognitive Vision Lab that focuses on computer vision and cognitive systems, in particular the learning of object representations in the context of grasping.

*Photo of the author(s)

Click here to download high resolution image