Neurocomputing 184 (2016) 3–12

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A comparison of feature detectors and descriptors for object class matching

Antti Hietanen [a], Jukka Lankinen [a], Joni-Kristian Kämäräinen [a,*], Anders Glent Buch [b], Norbert Krüger [b]

[a] Department of Signal Processing, Tampere University of Technology P.O. Box 553, FI-33101 Tampere, Finland
[b] Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Denmark

ARTICLE INFO

Article history:
Received 29 December 2014
Received in revised form
16 July 2015
Accepted 4 August 2015
Available online 2 December 2015

Keywords:
Local descriptor
Local detector
Interest point
SIFT
SURF
BRIEF

ABSTRACT

Solid protocols to benchmark local feature detectors and descriptors were introduced by Mikolajczyk et al. [1,2]. The detectors and the 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 collapses; (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.

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 [1] and Mikolajczyk et al. [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.

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] and LIOP [11]. 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 [12,13], 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 repeatability 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:

* Corresponding author at: Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland. Tel.: +358 50 300 1851.
E-mail address: joni.kamarainen@tut.fi (J.-K. Kämäräinen).

http://dx.doi.org/10.1016/j.neucom.2015.08.106
0925-2312/

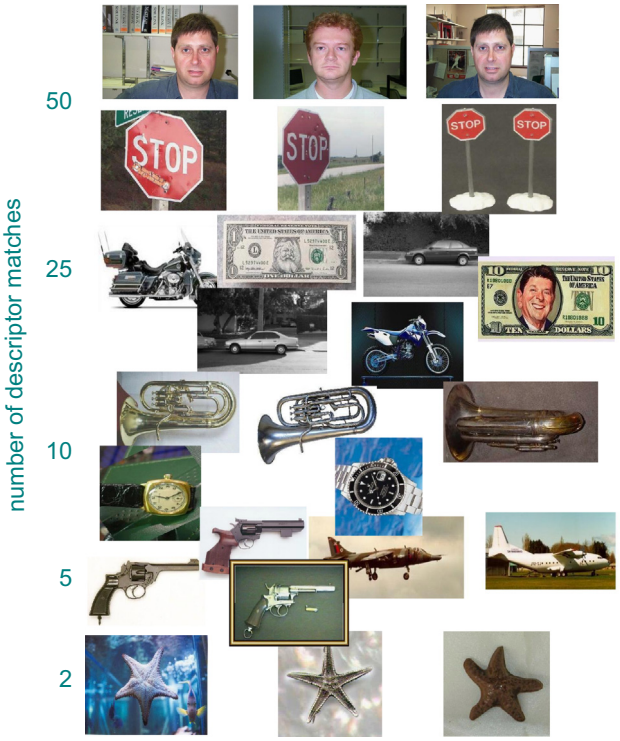

Fig. 1. Numbers of descriptor matches between two random class examples.

- 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.[1] 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. [14], but its main use is in visualization. More quantitative evaluations were reported by Zhang et al. [12] and Mikolajczyk et al. [13], but these were tied to a single methodology, the visual Bag-of-Words (BoW) [15,16]. 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.

[1] https://bitbucket.org/kamarain/descriptor_vocbenchmark

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., [17]), the descriptors (parts) should match despite substantial variance in their visual appearance.

2.1. Data

The experiments were conducted with the Caltech-101 [18] 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 [19]. The foreground masks were used to remove features detected in the background (Fig. 2). Affine correspondence between category examples was established by manually annotating 5–12 landmarks per category and estimating the pair-wise image transformations using the direct linear transform [20] 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 [21] and the recently proposed detectors: BRIEF [10], BRISK [9], ORB [8] and FREAK [7]. The preliminary detectors were

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 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 [22]) 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