J. Vis. Commun. Image R. 47 (2017) 62–72

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Performance evaluation of local descriptors for maximally stable extremal regions ☆

Man Hee Lee [a], In Kyu Park [b,*]

[a] Electronics and Telecommunication Research Institutes, Daejeon 34129, Republic of Korea
[b] Inha University, Incheon 22212, Republic of Korea

ARTICLE INFO

Article history:
Received 12 July 2016
Revised 17 April 2017
Accepted 17 May 2017
Available online 22 May 2017

Keywords:
Local feature
Feature descriptor
Performance evaluation
Affine-invariant region detectors
MSER

ABSTRACT

Visual feature descriptors are widely used in most computer vision applications. Over the past several decades, local feature descriptors that are robust to challenging environments have been proposed. Because their characteristics differ according to the imaging condition, it is necessary to compare their performance consistently. However, no pertinent research has attempted to establish a benchmark for performance evaluation, especially for affine region detectors, which are mainly used in object classification and recognition. This paper presents an intensive and informative performance evaluation of local descriptors for the state-of-the-art affine-invariant region detectors, *i.e.*, maximally stable extremal region detectors. We evaluate patch-based and binary descriptors, including SIFT, SURF, BRIEF, FREAK, the shape descriptor, LIOP, DAISY, GSURF, RFDg, and CNN descriptors. The experimental results reveal the relative performance and characteristics of each descriptor.

1. Introduction

Local feature detection and description are key ingredients in major computer vision tasks such as visual simultaneous localization and mapping (SLAM) [1], structure from motion [2], object retrieval [3], emotion prediction [4], and scene classification [5]. Several local feature detectors and descriptors have been proposed in the last two decades. For example, scale-invariant feature transform (SIFT) [6] and speeded-up robust features (SURF) [7] are conventionally used in a broad range of applications.

Each local detector has various invariance levels. A rigorous survey of the performance evaluation of local descriptors can be found in Mikolajczyk and Schmid's work [8]. Although local detectors such as Harris Laplace [9], SIFT, and SURF cope well with scale and rotation changes as well as photometric variation, they fail under significant viewpoint changes. In contrast, affine region detectors are more convenient for handling higher levels of invariance caused by affine transformation. Maximally stable extremal regions (MSER) [10], Harris Affine [9], and Hessian Affine [11] are a few notable examples of affine region detectors. These detectors extract the interest points with support regions that adapt to the geometric transformation of the image. Conventional affine region detectors do not have their own inherent descriptors. Although any major descriptor can be employed to represent the extracted regions, the choice of the proper descriptor remains nontrivial. Note that customized descriptors for describing affine-invariant regions, *e.g.*, the shape descriptor, have also been introduced [12].

In this study, we evaluate the performance of local descriptors for affine-invariant region features. To the best of our knowledge, no previous study has addressed this problem. By employing the MSER detector for the affine-invariant region detection, we compare both long-established and recently proposed local descriptors, including SIFT, SURF, local intensity order pattern (LIOP) [25], binary robust independent elementary features (BRIEF) [26], fast retina keypoint (FREAK) [27], the shape descriptor [12], DAISY [28], Gauge-SURF (GSURF) [29], the Gaussian receptive fields descriptor (RFDg) [30], and convolutional neural network (CNN) descriptor [31]. The performance of these descriptors is evaluated under different zoom levels, rotation, large viewpoint change, object deformation, and large depth variation. The preliminary result of this paper was presented in [32]. In this extended version, we explore more descriptors and provide an extensive evaluation with additional variations in imaging condition.

The rest of the paper is organized as follows. In Section 2, the existing performance evaluations are introduced briefly. Section 3 describes the evaluation framework and criteria and a brief summary of the individual detector and descriptors being compared. The experimental results and discussion are presented in Section 4. Finally, we provide conclusive remarks in Section 5.

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.

E-mail addresses: mheelee@etri.re.kr (M.H. Lee), pik@inha.ac.kr (I.K. Park).

http://dx.doi.org/10.1016/j.jvcir.2017.05.008
1047-3203/© 2017 Elsevier Inc. All rights reserved.

Table 1
Previous works on performance evaluation of feature detectors/descriptors.

Author	Type	Environment	Best result
Mikolajczyk [8]	Local descriptor	Geometric + photometric transform	GLOH, SIFT
Miksik [13]	Local descriptor	Accuracy and speed	LIOP, BRIEF
Kaneva [14]	Local descriptor	Viewpoint and illumination change	DAISY
Heinly [15]	Binary descriptor	Geometric + photometric transform	BRIEF
Restrepo [16]	Shape descriptor	Object classification	FPFH
Moreels [17]	Detector + descriptor	3D object	Hessian-affine + SIFT
Gil [18]	Detector + descriptor	Visual SLAM	GLOH, SURF
Dahl [19]	Detector + descriptor	Multi-view dataset	MSER + SIFT
Gauglitz [20]	Detector + descriptor	Visual tracking	Fast Hessian + SIFT
Mikolajczyk [11]	Affine region detector	Geometric + photometric transform	MSER
Haja [21]	Region detector	Texture + structure	MSER
Schmid [22]	Local detector	Geometric + photometric transform	Harris
Dickscheid [23]	Local detector	Image coding	MSER
Canclini [24]	Local detector	Image retrieval	BRISK

2. Related work

Table 1 presents a complete list of the previous performance evaluations of feature detectors and descriptors. Mikolazcyk and Schmidt [8] evaluated the performance of local feature descriptors under various geometric and photometric transformations and has conducted the most exhaustive study so far. In addition, the gradient location and orientation histogram (GLOH) descriptor was proposed as an extension of the SIFT descriptor by applying a log-polar grid for gradient quantization. It was shown that the combination of GLOH and SIFT outperformed the other descriptors under rotation, zoom, blur, image compression, viewpoint, and illumination changes. Moreels and Perona [17] investigated the performance of the popular detectors and descriptors for 3D objects. They generated a database of 144 objects with viewpoint and illumination changes. An evaluation of several combinations of feature detectors and descriptors revealed that the combination of the Hessian-Affine detector and SIFT descriptor outperformed the other detectors and descriptors under viewpoint and illumination changes in a 3D setup. Gil et al. [18] compared the behavior of different feature detectors and descriptors for visual SLAM. They evaluated the repeatability of the detectors as well as the invariance and distinctiveness of the descriptors. In their experiments, GLOH and SURF were the most suitable for visual SLAM. Dahl et al. [19] investigated feature detector and descriptor combinations for a multiview dataset. The MSER and difference-of-Gaussian (DoG) detectors together with the SIFT descriptor provided the best results in their experiment. Miksik and Mickolaczyk [13] assessed the trade-off between speed and accuracy for local descriptors. They evaluated the performance of several binary descriptors and local intensity order descriptors. Their results showed that the binary descriptors outperformed other descriptors in time-constrained applications with low memory requirements. Kaneva et al. [14] compared the performance of local descriptors in terms of viewpoint and illumination changes. They simulated a controlled condition with photorealistic synthetic scenes and concluded that the DAISY descriptor worked best under these changes.

The performance of local shape descriptors for object classification was evaluated by Restrepo and Mundy [16]. The local shape descriptors were extracted from the probabilistic volumetric model. These researchers compared several shape descriptors to classify object categories using the Bag of Words model from large-scale urban scenes. In their experiments, the fast point feature histogram (FPFH) descriptor [33] showed good performance. Gauglitz et al. [20] evaluated the performance of feature detectors and descriptors for visual camera tracking. In their experiments, the center-oriented detectors provided the highest repeatability. Schmid et al. [22] evaluated the performance of low-level feature detectors. They introduced two evaluation criteria: repeatability and information content. Repeatability compares the geometrical stability of features under different transformations, whereas information content measures the distinctiveness of features. They concluded that the improved version of Harris outperformed the other detectors studied. Dickscheid et al. [23] measured the completeness of local features for image coding. They proposed a qualitative metric for evaluating the completeness of feature detection using feature density and entropy density. In their experiment, the MSER detector achieved the best performance. Canclini et al. [24] evaluated the performance of feature detectors and descriptors for image retrieval applications. They compared several low-complexity feature detectors and descriptors, concluding that binary descriptors outperform non-binary descriptors in terms of matching accuracy and computational complexity.

Although those studies provide a rich understanding of the performance of local descriptors, they do not consider newly proposed descriptors. At present, computer vision applications continuously require the performance evaluation of contemporary state-of-the-art algorithms, which is the main motivation of this paper. In this paper, we narrow our focus to the performance evaluation of descriptors combined with state-of-the-art affine-invariant region detection, *i.e.*, the MSER detector. This combination has not yet been addressed in the literature. In this study, we include recent descriptors that were absent in previous studies.

3. Performance evaluation framework

3.1. Affine invariant region detector

Affine-invariant detectors extract regional features adjusted to the geometric transformation of the image. Several techniques have been proposed, such as MSER, Harris Affine, and Hessian Affine. The Harris Affine detector [9] detects corner-like features and is based on the Harris-Corner. On the other hand, the Hessian Affine detector [11] is based on a Hessian matrix that responds strongly to blob-like features. Both Harris and Hessian Affines utilize the Laplacian of Gaussian extrema for scale selection. They iteratively estimate the elliptical affine regions of detected features using a second-order moment matrix.

The MSER detector [10] extracts regional features that are stable under geometric and photometric transformation. Each detected feature is a connected region that is either darker or brighter than its surroundings, as shown in Fig. 1. MSER has several desirable properties:

- Invariance to affine transformation and image intensity change.
- Preservation of adjacency between neighboring components for continuous geometric transformation.