Signal Processing: Image Communication 59 (2017) 150–167

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

A comprehensive evaluation of local detectors and descriptors

Song Wu[a], Ard Oerlemans[b], Erwin M. Bakker[c], Michael S. Lew[c,*]

[a] College of Computer and Information Science, Southwest University, Chongqing, China
[b] VDG Security BV, Zoetermeer, The Netherlands
[c] LIACS Media Lab, Leiden University, Niels Bohrweg 1, Leiden, The Netherlands

ARTICLE INFO

Keywords:
Local detector
Local descriptor
Visual words model
Fully affine space
Evaluation

ABSTRACT

As local detectors and descriptors can find and represent distinctive keypoints in an image, various types of keypoints detection and description methods have been proposed. Each method has particular advantages and limitations and may be appropriate in different contexts. In this paper, we evaluate the performance of a wide set of local detectors and descriptors. First, we compare diverse local detectors with regard to the repeatability, and local descriptors in terms of the recall and precision. Next, we apply the visual words model constructed from the local descriptors with real values and binary string to large scale image search. The evaluation results reveal some strengths and weaknesses of the recent binary string descriptors compared with the notable real valued descriptors. Finally, we integrate the local detectors and descriptors with the framework of fully affine space and evaluate their performance under major viewpoint transformations. The presented comparative experimental studies can support researchers in choosing an appropriate local detector and descriptor for their specific computer vision applications.

1. Introduction

Local detectors and descriptors which can locate and describe meaningful, stable, and representative keypoints in an image have become prevalent in diverse areas in computer vision, such as object and scene recognition [1,2], 3D object reconstruction [3], visual tracking [4,5] and multimedia information retrieval [6–14]. Most of the local keypoints algorithms contain two parts: a detector and a descriptor. The detector locates a set of distinctive points which can be invariant to various transformations (e.g. scaling, translation, viewpoint changes), meanwhile the descriptor encodes the important information from the local patch centered on the keypoint into a feature vector, which allows to reliably match correspondences across different transformations of the same object or the same scene.

Typically, object recognition, 3D reconstruction and visual tracking mainly rely on the correctly matched correspondences between two compared images. These applications start by extracting local descriptors from each image and insert the obtained local descriptors into an index space for efficient correspondence matching. The RANSAC algorithm [15] is further adopted to eliminate outlier matches and to estimate the homography between the compared images. Therefore, a local detector with high repeatability and a local descriptor with discriminatory power is required for these applications.

Moreover, empirical experiments conducted over the past decade have demonstrated that one of the most popular and successful approaches towards similar image and visual concept analysis is to use local keypoints combined with visual words and approximate nearest neighbors (ANN) search [16]. This is mainly the result of local descriptors that are distinctive under various geometric transformations and also the introduction of the visual words model, which significantly improved the search efficiency and the adaptability to a particular image dataset. Current visual words systems are predominantly built using SIFT [17] and SURF [18] whose descriptor type is a real value. In contrast to the real value descriptors, binary string descriptors were proposed in order to generate the feature descriptors more efficiently (i.e. BRIEF [19], ORB [20], BRISK [21], FREAK [22], BinBoost [23] and LATCH [24]). Another goal of this paper is to give insights into the performance and requirements of these descriptors for large scale image search.

However, accurate correspondence matching under large viewpoint changes is still a major challenge, because greater image viewpoint

[*] Corresponding author.
E-mail addresses: songwuswu@swu.edu.cn (S. Wu), a.oerlemans@vdgsecurity.com (A. Oerlemans), e.m.bakker@liacs.leidenuniv.nl (E.M. Bakker), m.s.lew@liacs.leidenuniv.nl (M.S. Lew).

http://dx.doi.org/10.1016/j.image.2017.06.010
Received 5 January 2017; Received in revised form 1 June 2017; Accepted 23 June 2017
Available online 1 July 2017
0923-5965/© 2017 Published by Elsevier B.V.

transformations result in a significant decrease of saliency and repeatability of keypoints. Yu et al. [25] proposed to use the framework of fully affine space to overcome this issue. The basic idea behind the framework of fully affine space is that the projective transformation induced by camera motion around a smooth surface can be approximated by an affine transformation. A notable method is ASIFT which generates all image views in the whole affine space and extracts SIFT local features in these synthetic images to increase the matching precision. As the high dimensionality of the SIFT descriptor leads to a high computational complexity in the framework of fully affine space, we combine the recent lower computational complexity local detectors and descriptors with the framework of fully affine space and evaluate their performance under the extreme viewpoint changes.

Several related reviews present the performance evaluation of various local detectors and descriptors. In contrast to these related local detector and descriptor reviews [5,26–32], our work mainly focus on evaluating the performance of visual words representation conducted on local descriptors for large scale image search, as well as testing the viewpoint invariance of each local detector and descriptor in the fully affine space.

The main contributions of this paper are summarized as follows:

First, the repeatability performance and the computational cost of each local detector are presented. Additionally, the efficiency and accuracy of both the real valued descriptors and binary string descriptors in terms of recall and precision on two benchmark datasets are evaluated.

Second, the visual words constructed from real value descriptors and binary string descriptors are evaluated for the application of large scale image search.

Third, we calculate the accuracy and time complexity of each local detector and descriptor in the framework of fully affine space such that researchers could make a trade-off between precision and efficiency under extreme viewpoint changes.

The rest of the paper is organized as follows: In Section 2, we present the background as well as an overview of recent local detectors and descriptors. In Section 3, we present the generation of real value type and binary string type visual words. In Section 4, we describe the details of the fully affine space framework. The details of experimental setup are described in Section 5. The evaluation results and discussions are given in Section 6, and conclusions are given in Section 7.

2. Overview of local detectors and descriptors

Early research about keypoints algorithms mainly focused on finding high variance or corner points in the image. One of the first local detectors was developed by Moravec [33] and it is defined according to the average intensity changes in different directions within the local region around a point. The Harris corner detector [34] defines a corner structure point, if its second-moment matrix has two large eigenvalues. The similar Hessian corner detector [35] determines a corner point in the image, if it is the local extrema of the Hessian matrix determinant. As both the Harris and Hessian detectors find the corner points at a fixed scale, the Harris–Laplacian and Hessian–Laplacian [36,37] are designed to be scale invariant. Harris–Laplacian and Hessian–Laplacian locate corner candidates on each level of the scale space. Those points for which the Laplacian simultaneously attains local extrema over scales are selected as corner points.

Since conventional corner point detectors are only invariant to scale, translation, and noise, affine covariant region detectors were developed to reduce the influence of viewpoint changes. The Harris-Affine detector and the Hessian-Affine detector [38] find the initial candidate points by using the Harris–Laplacian corner detector and Hessian–Laplacian corner detector, respectively, and then fit an elliptical region to each point via the second moment matrix of the intensity gradient. MSER [39] computes the connected binary regions through a large set of multiple thresholds, and the selected regions are those that maintain unchanged shapes over these thresholds. As edges are typically rather stable structures that can be detected over a range of image changes, EBR [40] starts by detecting corner points in an image and identifies the affine covariant region of each point by exploiting the edge information present nearby. IBR [41] detects intensity extrema at multiple scales and captures the intensity pattern along rays emanating from each extremum to define a region of arbitrary shape. The region of IBR is delineated by the image points defined over these rays where the intensity suddenly increases or decreases, and then uses an ellipse to fit the region. However, the operation of elliptical region fitting in the affine covariant detector could result in partial information loss.

Recent keypoints methods focus on the repeatability and precision of the local detector, as well as the distinctiveness, computational efficiency and low memory requirement of the local descriptor. This is mainly achieved by four procedures: the first step is to establish the scale space and find the extrema across all scales to achieve scale invariance. The second step is to determine the locations of the extrema and to define a local region for each according to the scale information. Then, each defined region is normalized and assigned a domain orientation to be rotation invariant. Finally, the region content is rotated based on the calculated orientation, after which, the discriminative information in the rotated region is encoded into a local descriptor.

The most representative local keypoints method is SIFT. Meanwhile, some variants of SIFT are proposed with the aim to increase the discrimination of the SIFT descriptor. PCA-SIFT [42] utilizes PCA to reduce the dimension of the original SIFT descriptor to further speed up the process of local descriptor matching. Color-SIFT [43] takes the color gradients, rather than intensity gradients in the local region around the keypoint to generate the feature. Rank-SIFT [44] adopts a data-driven approach to learn a ranking function to sort the keypoints such that the unstable points can be discarded. Root-SIFT [45] adds a square root operation to the normalized SIFT features and uses the Hellinger kernel to increase the matching accuracy. DSP-SIFT [46] generates the descriptor through pooling the gradient histogram across different domain sizes of each keypoint into a feature and it even outperforms the high level convolutional neural network feature [47]. Affine-SIFT (ASIFT) [25] is proposed with the aim to be perspective invariant and it does this by simulating images under various views to cover the whole affine space and extracting SIFT descriptors in all these simulated images for matching. Different from these variants of SIFT, other approaches target on improving the efficiency of scale space establishment or accuracy of keypoints localization. For example, the SURF detector makes use of a box-filter and the integral image to speed up the scale space build. The ORB and BRISK detectors use a Gaussian image pyramid to efficiently establish the scale space. As the construction of scale space by linear multi-scale Gaussian pyramids easily results in the blurring and the loss of boundary details, KAZE [48] combines a nonlinear scale space with Additive Operator Splitting (AOS) and special conductance diffusion to reduce noise while retaining the object boundary structure.

In order to meet the requirements of real time systems and devices with limited computational and storage resources, binary string local descriptors were recently introduced. Binary string representations make use of a pixel-pair intensity comparison to generate the binary code. The resulting binary code holds some significant advantages: first, the operation of intensity comparison is fast, the memory requirement of binary codes is low and matching binary codes via the hamming distance is much faster than the Euclidean metric.

Table 1 gives an overview of all the evaluated local detectors and descriptors in the experiments section. The details of each evaluated local detector and descriptors are also given as follows.

2.1. SIFT (detector/descriptor)

SIFT proposed by Low [17] is the most popular keypoint approach. The implementation of SIFT begins by building the Gaussian scale space which approximates the Laplacian-of-Gaussian function by the computationally efficient Difference-of-Gaussian function. It searches extrema over all scales to identify the potential keypoints. Since the