Computer Vision and Image Understanding 162 (2017) 71–86

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

mdBRIEF - a fast online-adaptable, distorted binary descriptor for real-time applications using calibrated wide-angle or fisheye cameras

Steffen Urban*, Martin Weinmann, Stefan Hinz

Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Englerstr. 7, 76131 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 29 November 2016
Revised 22 August 2017
Accepted 28 August 2017
Available online 5 September 2017

Keywords:
Visual features
Feature detection
Feature description
Binary descriptors
Hamming distance

ABSTRACT

Fast binary descriptors build the core for many vision based applications with real-time demands like object detection, visual odometry or SLAM. Commonly it is assumed, that the acquired images and thus the patches extracted around keypoints originate from a perspective projection ignoring image distortion or completely different types of projections such as omnidirectional or fisheye. Usually the deviations from a perfect perspective projection are corrected by using standard undistortion models. The latter, however, introduce artifacts if the camera's field-of-view gets larger. In addition, many applications (e.g. monocular SLAM) require only undistorted points and holistic undistortion of every image for descriptor extraction could be eluded. In this paper, we propose a distorted and masked version of the BRIEF descriptor for calibrated cameras, called dBRIEF and mdBRIEF respectively. Instead of correcting the distortion holistically, we distort the binary tests and thus adapt the descriptor to different image regions. The implementation of the proposed method along with evaluation scripts can be found online at https://github.com/urbste/mdBRIEF.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

The detection and description of salient image features is at the core of most computer vision or photogrammetric processing chains. Exemplary applications include but are not limited to image retrieval (Qin et al., 2016), 3D reconstruction (Rumpler et al., 2016), localization (Gálvez-López and Tardós, 2012) and scan registration (Urban and Weinmann, 2015). Addressing different requirements, numerous methods for feature extraction have been developed over the past decades. The proposed features can be grouped into different categories and typically reveal both advantages and disadvantages, depending on the task, the application or the platform.

The extraction of salient image features commonly consists of three main steps. The first step focuses on *feature detection*, whereby the image is searched for keypoints. Such keypoints are typically represented by either corners or blob-/region-like structures, depending on the used detector. A desirable property of such detectors is that they should be able to repeatably detect the same keypoint over viewpoint changes. In addition, a robust estimation of the orientation as well as the scale at which the keypoint was detected is necessary for some applications. Once features have been detected, the second step focuses on *feature description* aiming at a robust description of the image content surrounding a keypoint. Thereby, the description is usually represented as a vector consisting either of floating point values or binary strings, depending on the used descriptor. If the keypoint detector provides rotation or scale information, the descriptor can be made invariant to such viewpoint changes. For many tasks including egomotion estimation, this property is of fundamental necessity as the movement of the camera introduces a continuous change of viewpoint. Finally, the third step focuses on *feature matching* where descriptors are matched by comparing their vector representations. A match is found if a specific criterion is satisfied, e.g. based on the distance between descriptors or based on heuristics like the distance ratio test (Lowe, 2004). Depending on the vector type, either the L2-Norm (floating point) or the Hamming distance (binary strings) can be computed. The latter is more efficient, as the instruction sets (SSE, AVX) of modern processors allow for parallel bit counting. In addition, different matching strategies exist such as matching each descriptor from one image to each descriptor from another image, or applying fast approximate nearest neighbor algorithms (Muja and Lowe, 2009). At this point it is important to emphasize that there does not exist one perfect feature for all tasks. Some descriptor might outperform all others in terms of recognition rate on one dataset, but might be impractical for real-time applications due to a high computational burden.

* Corresponding author.
E-mail addresses: steffen.urban@kit.edu (S. Urban), martin.weinmann@kit.edu (M. Weinmann), stefan.hinz@kit.edu (S. Hinz).
URL: http://www.ipf.kit.edu (S. Urban)

http://dx.doi.org/10.1016/j.cviu.2017.08.011
1077-3142/© 2017 Elsevier Inc. All rights reserved.

In this paper, we focus on feature description for online omnidirectional image matching, a topic which is of particular interest for applications such as SLAM and visual odometry. To handle significant image distortions as given for wide-angle, fisheye or omnidirectional images and thereby retain computational efficiency, we combine the advantages of the existing BRIEF, ORB and BOLD descriptors (Balntas et al., 2015; Calonder et al., 2010; Rublee et al., 2011) by presenting a distorted and masked version of the BRIEF descriptor for calibrated cameras, called *dBRIEF* and *mdBRIEF* respectively. Instead of correcting the distortion holistically, we distort the binary tests and thus adapt the descriptor to different image regions. In summary, the main contributions of our paper are:

- A distorted version of the BRIEF descriptor that adapts to local image distortions.
- An offline and online learned version of the distorted descriptor.
- A thorough analysis on the impact of distortion on matching performance.
- A comparison to various state-of-the-art descriptors.

The implementation of the proposed dBRIEF and mdBRIEF descriptors as well as evaluation scripts can be found online at https://github.com/urbste/mdBRIEF.

The paper is organized as follows. First, we briefly summarize related work on feature extraction from imagery in Section 2. Thereby, we specifically consider approaches that are of particular interest in the context of our work, i.e. in the sense of developing a real-time capable descriptor for highly distorted images. Subsequently, in Section 3, we extend and adopt recent descriptors for the use in online omnidirectional image matching and especially for applications such as SLAM and visual odometry. The proposed distorted versions of the BRIEF descriptor – which we refer to as dBRIEF – are tested in Section 4. In Section 5, we adapt the dBRIEF descriptor by online mask learning which yields the mdBRIEF descriptor with increased efficiency regarding feature matching. To demonstrate the performance of the proposed dBRIEF and mdBRIEF descriptors, we present experimental results achieved on real data in Section 6. Finally, we provide concluding remarks in Section 7.

2. Related work

This section gives an overview of the state-of-the-art for keypoint detectors and descriptors. As this is a vast topic, we can only cover a limited amount of related work. Instead, this section is supposed to analyze the currently available methods and help to find, adapt and tune a good feature detector-descriptor combination to the given task of egomotion estimation. For more comprehensive overviews as well as evaluations of recent research and early developments, the reader is referred to surveys of Weinmann (2013), Fan et al. (2015), Miksik and Mikolajczyk (2012), Gauglitz et al. (2011), Mikolajczyk and Schmid (2005), Heinly et al. (2012) and Tuytelaars and Mikolajczyk (2008).

To narrow down the state-of-the-art relevant to the task addressed in this work, we first analyze the requirements that a detector-descriptor combination for egomotion estimation has to meet. First of all, the 3-stage process of detection, description and matching has to be performed in real-time (we consider at least 25 Hz). In addition, many descriptors are supposed to be used in place recognition and re-localization tasks. Thus, the storage requirements are not negligible as a database of descriptors has to be built. Apart from the processing speed, the detector-descriptor combination should be invariant (to some extent) to viewpoint and illumination changes, as the motion of a camera can be unconstrained in terms of rotation and translation. Moreover, the descriptors should be invariant under image distortion or fisheye projections. As we will see, the latter is a requirement rarely addressed. As today's camera systems are supposed to operate in a changing environment and will visit the same scene more than once, an online adaption of the descriptors that are stored in the database would also be favorable.

2.1. Keypoint detectors

In general, keypoint detectors can be grouped into two categories, i.e. corner and blob-/region-like detectors (Table 1).

Corners. Two of the first and still widely used corner detectors are the Harris corner detector (Harris and Stephens, 1988) and the "Good Features to Track" detector (Shi and Tomasi, 1994). In both methods, a "corner score" is derived from the second-order moment matrix which involves computing the image derivatives. To avoid such costly computations, the FAST detector (Rosten and Drummond, 2006; Rosten et al., 2010) uses simple gray value comparisons on discretized circle segments around a candidate pixel to decide whether a pixel is classified as a corner. To speed up the process, machine learning techniques are involved (a ternary decision tree) to reduce the number of necessary pixel comparisons to a minimum. This decision tree should be relearned for new environments or scene structures, which has been considered with the AGAST detector (Mair et al., 2010). To additionally address changes in image scale and orientation, the ORB detector (Rublee et al., 2011) has been proposed which builds an image pyramid, extracts FAST features on every level and estimates a salient orientation per feature. The BRISK detector (Leutenegger et al., 2011) tries to improve upon ORB by extracting AGAST corners on different levels of the image pyramid and scale interpolation between octaves.

Blobs/Regions. Blobs are usually detected as the local extrema of different filter responses. The SIFT detector (Lowe, 2004) generates a Gaussian scale space and then detects extrema in differences of the Gaussian filtered images. To speed up the computation, the SURF detector (Bay et al., 2006) approximates the Gaussian derivatives by box filters. In addition, both detectors estimate a salient orientation. The authors of the KAZE (Alcantarilla et al., 2012) detector as well as its accelerated pendant AKAZE (Alcantarilla et al., 2013) argue that a linear Gaussian scale space smooths noise and object details to the same degree. Hence, they build a non-linear scale space using non-linear diffusion filtering to obtain keypoints that exhibit a much higher repeatability. Stable image regions are found by the MSER detector (Matas et al., 2004) that identifies sets of connected pixels above a threshold.

So far, none of the presented detectors considers image distortion or different projections. Both introduce deformations to corners and blobs that will decrease the repeatability of keypoints. In theory, the image could be warped to match a perfect planar perspective projection. This however introduces severe artifacts and is limited to images with less than a hemispherical view of the scene and should be avoided (Daniilidis et al., 2002). Thus, all image processing operations have to be carried out directly in the distorted image. To detect repeatable keypoints in omnidirectional images, invariant to scale and rotation, the construction of the scale space (Bülow, 2004; Hansen et al., 2008; Puig and Guerrero, 2011) as well as the image gradient computation have to be adjusted (Furnari et al., 2015). Prominent enhancements of the SIFT detector are pSIFT (Hansen et al., 2008) and sRD-SIFT (Lourenço et al., 2012) which excel their "standard" pendants regarding matching performance. However, the adapted construction of scale spaces on the sphere increases their extraction time significantly and makes them even less suitable for real-time applications. An enhancement of the FAST detector regarding image distortion would require adjusting (distorting) the discretized circle

Download English Version:

https://daneshyari.com/en/article/4968697

Download Persian Version:

https://daneshyari.com/article/4968697

Daneshyari.com