EVALUATION ON STRUCTURE FROM MOTION USING BINARY ROBUST INVARIANT SCALABLE KEYPOINTS

Min-Lung Cheng, Masashi Matsuoka

Tokyo Tech, 4259-G3-2, Nagatsuta-cho, Midori-ku, Yokohama, 226-8502, Japan
Email: cheng.m.ab@m.titech.ac.jp; matsuoka.m.ab@m.titech.ac.jp

KEY WORDS: SfM, pose estimation, 3D reconstruction, BRISK

ABSTRACT: Camera poses recovery and three-dimensional (3D) reconstruction from a succession of optical imagery by structure from motion (SfM) has been an interesting study for years. In order to retrieve the camera positions and angular information, common feature points in any stereo pair are necessary, even to the later 3D reconstruction. Under the category of local-based features, Scale Invariant Feature Transform (SIFT) is a representative keypoint, which has been broadly utilized in SfM. However, to match such vector-based features requires computational time to achieve the work. To improve that demerit, this paper tries to replace SIFT keypoints with Binary Robust Invariant Scalable Keypoints (BRISK), which is classified as binary-based features, to perform SfM since BRISK matching is claimed consuming less computational power.

This paper addresses a standard procedure of SfM with BRISK features to solve the camera poses and achieve sparse 3D reconstruction by using a sequence of close-range images. Relative orientations and reconstructed 3D objects points can be acquired by processing several independent stereo models in advance. Common 3D points within those independent models can then be exploited to joint each model. Therefore, a complete flight path as well as the 3D reconstruction can be achieved. Optimization by bundle adjustment can improve the reliability of the acquired relative camera orientations and the 3D object points as a last step. The experiment also compares the computational costs and the accuracy of 3D reconstruction by SIFT and BRISK features, respectively. It is shown that using BRISK features has higher computational efficiency as twice more than SIFT keypoints in the 3D reconstruction.

1. INTRODUCTION

3D reconstruction has been one of the most significant topics when extracting spatial information from optical imagery for years. This development not only aids in realizing the earth from a border viewpoint, but also changes the way to understand the activities on the land surface. It also brings a variety of spatial applications to solve environmental problems such as disaster analysis and mitigation, escaping strategies conduction and simulation in an area, and future smart city planning. Therefore, a lot of environment-related works can be visualized in the computer system, and then appropriate solutions can be figured out beforehand. For instance, a 3D cyber city can simulate the flooding status within a region when there is a heavy rainfall, and thus an escaping strategy can be proposed in advance. Such merits of 3D reconstruction might reduce the cost needed and improve the efficiency when coping with environmental problems.

In the past decades, local descriptors in an optical image have played a role in feature detection, matching, and 3D reconstruction. Compared to window-based image features which mainly rely on gradient information, local descriptors contain more reliable traits such as scale and rotation invariances. These characteristics make the feature representations more stable and for image matching. A broadly used image feature (keypoint) is SIFT (Lowe, 2004) owning to its well-organized scale-invariant and rotation-resistant manners. For most SfM processing, it can be said as the most powerful local descriptor in the algorithm. But there is also a disadvantage of this vector-based keypoint. The computational efficiency is often slow because of the usage of Euclidean distance to measure the similarity between two feature descriptors. Although later improved versions of Speed-Up Robust Features (SURF) (Bay et al., 2008) and PCA-SIFT (Ke and Sukthankar, 2004) are developed, a balance between data processing time and accuracy of the outcome is still hard to achieve.

An alternative image keypoint known as binary-based features has competitive potential with vector-based features in terms of computational efficiency. Binary Robust Invariant Elementary Feature (BRIEF) (Calonder et al., 2010) and Features from Accelerated Segment Test (FAST) (Rosten et al., 2006) are pioneer keypoints in this category. A later improvement toward rotation invariance appearing as Orientated FAST and Rotated BRIEF (ORB) (Bradski et al., 2011) enhances the binary-based feature matching. However, a weak point is still remained that the scale of a keypoint can't be determined when using ORB features. Toward this demerit of ORB features, Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011) is an advanced replacement to conquer the weakness of both rotation and scale invariances for binary-based features. BRISK has been proven useful and

powerful in the aspect of matching efficiency and accuracy comparing to SIFT for image matching. Thus, this paper gives initial investigations on camera pose recovery and 3D reconstruction via replacing BRISK with SIFT for SfM implementation. Experimental results interpret that the modified word flow has a superior performance than traditional approaches in terms of computational cost, while the accuracy of the recovered camera pose and 3D reconstruction is comparable.

2. METHODOLOGY

In order to recover the camera poses and reconstruct 3D points by using a sequence of images, this paper aims at replacing SIFT keypoints with BRISK features for fast data processing. With an input stereo pair of two images, BRISK features are detected and extracted. These binary-based BRISK features are then matched by the Hamming distance to determine a point in the master image and its correspondence in the slave image. After obtaining a group of matches in a stereo pair, SfM can utilize them to estimate the camera pose and reconstruct 3D object points. A proposed work frame is demonstrated as Figure 1, which interprets the strategies for evaluating the SfM technique via using BRISK features.

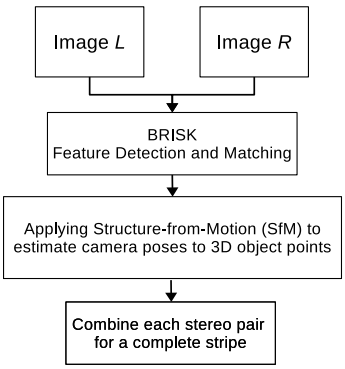


Figure 1. The proposed work flow of applying BRISK features for camera pose estimation and 3D reconstruction

2.1 Binary Robust Invariant Scalable Keypoints (BRISK)

Among the binary-based feature descriptors, BRISK is one of the most potential features against vector-based descriptors (e.g., SIFT and SURF) according to its rotation and scale invariants. To gain a BRISK feature, the corner detection algorithm of Features from Accelerate Segment Test (FAST) is carried out as the first step. The scale of a BRISK keypoint can be found by a scale-space pyramid and the FAST score, showing as Figure 2(a). After the scale factor is established, forming a string of binary-based descriptors to describe a BRISK feature can be achieved. The classic BRISK sampling pattern, displaying as Figure 2(b), is applied to generate 64-byte descriptors for each keypoint.

(a) (b)

Figure 2. The generation of a BRISK features and its sampling pattern (Leutenegger et al., 2011)

When using binary-based features for image matching, the Hamming distance is usually addressed to measure the similarity between two features. Since the 64-byte descriptors can be decomposed into 512-bit elements of 0 and 1, the Hamming distance is thus computed by counting the number of differences of two strings of BRISK descriptors. Figure 3 illustrates the mechanism of Hamming distance computation for feature similarity determination. A final decision can therefore be made by setting a threshold to judge whether two keypoints are resembling enough or not. Also, it has to be noted that a successful matched feature pair presents less Hamming distance because such as similarity measurement is built upon evaluating the quantity of differences of two image features.

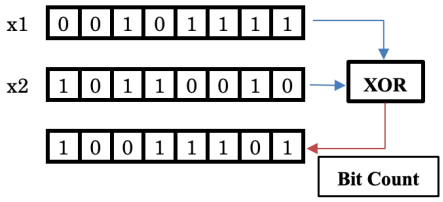


Figure 3. Hamming distance computation by comparing the differences between two strings of 0 and 1feature descriptors

2.2 Camera Pose Recovery and 3D Reconstruction

Structure from Motion (SfM) is a method that can recover the camera poses and reconstruct the 3D points without preliminary information (e.g., initial approximations for camera pose and 3D coordinates). By the matches of two images, this approach is able to estimate relative positions and rotations for the input data based on epipolar geometry and epipolar constraint. In this stage, the relationship of a stereo pair is established based on a fundamental matrix (or essential matrix). By this matrix, there are two sets of possible solutions for angular attitudes and positions, respectively.

To pick up the most probable combination of relative rotational and positional for the input two images, four groups of 3D points are firstly reconstructed by the four candidates of relative camera orientations. A second step is to compare which solution of camera pose has the greatest number of 3D points reconstructed in front of both two cameras. This procedure gives the most probable answer to determine the relative camera pose of a stereo model when exploiting 2D image features only. Consequently, this strategy sets the master photo as the origin and the approximates the rotational and positional information for the slave photo by means of relative camera pose. In addition, 3D points can be acquired simultaneously once the camera poses are determined.

Since SfM handles two images and produces an independent stereo model each time, connecting several stereo models is needed when there is a succession of images. A 3D conformal transformation is usually applied to link independent stereo models as there are more than three common object points in each model. The camera pose can also be added in order to enhance the geometry when performing this work. An alternative approach in this paper is relying on the same camera appearing in two stereo models, but their rotational angles and positions are different. By an aspect to unify the pose information from an independent model to another for the same camera, a transformation between two independent models can be found. However, a point should be noted that this method works only under the domain of an up-to-scale coordinate system or homogeneous space. For this reason, all elements, including camera poses and 3D object points, can be moved from one independent stereo model to another to joint several reconstructed independent 3D models.

3. EXPERIMENTAL RESULTS AND ANALYSIS

This paper exploits five close-range images to evaluate the performances between BRISK and SIFT for 3D reconstruction. Figure 4 demonstrates the sequence of images used in this experiment with detected BRISK and SIFT features. The projections of reconstructed 3D points are also displayed in the figure to investigate the reliability of the results. In the experimental outcomes, it is apparent that both BRISK and SIFT are probable to misestimate the camera poses, which lead to false 3D reconstruction several image pairs in Figure 4. By using 2D images features only to approximate the camera poses and reconstruct 3D points might be rely on the spatial distribution of the detected keypoints through an image. By projecting the reconstructed 3D points back to the 2D image using the estimated camera poses, the reprojection errors can be inspected to evaluated the reliability of SfM. In this experiment, both BRISK and SIFT cast good outcomes according to Figure 4 with average reprojection errors are less than 10 pixels.

In addition, the processing costs of BRISK and SIFT descriptor matching as well as the number of matches of the first stereo model in Figure 4 are expressed in Table 2. This example shows that BRISK matching (by FAST corner detection of threshold $t = 67$) outperforms SIFT matching in the perspective of computational cost and the quantity of matches. Although the entire computation time of BRISK features is comparable to SIFT keypoints, the produced quantity of 3D points of BRISK is nearly twice more than SIFT. Therefore, it is believed that BRISK features have a competitive potential than SIFT in optical image processing and spatial information extraction. Examples of the reconstructed 3D points by employing SIFT and BRISK feature descriptors are also displayed in Figure 5.

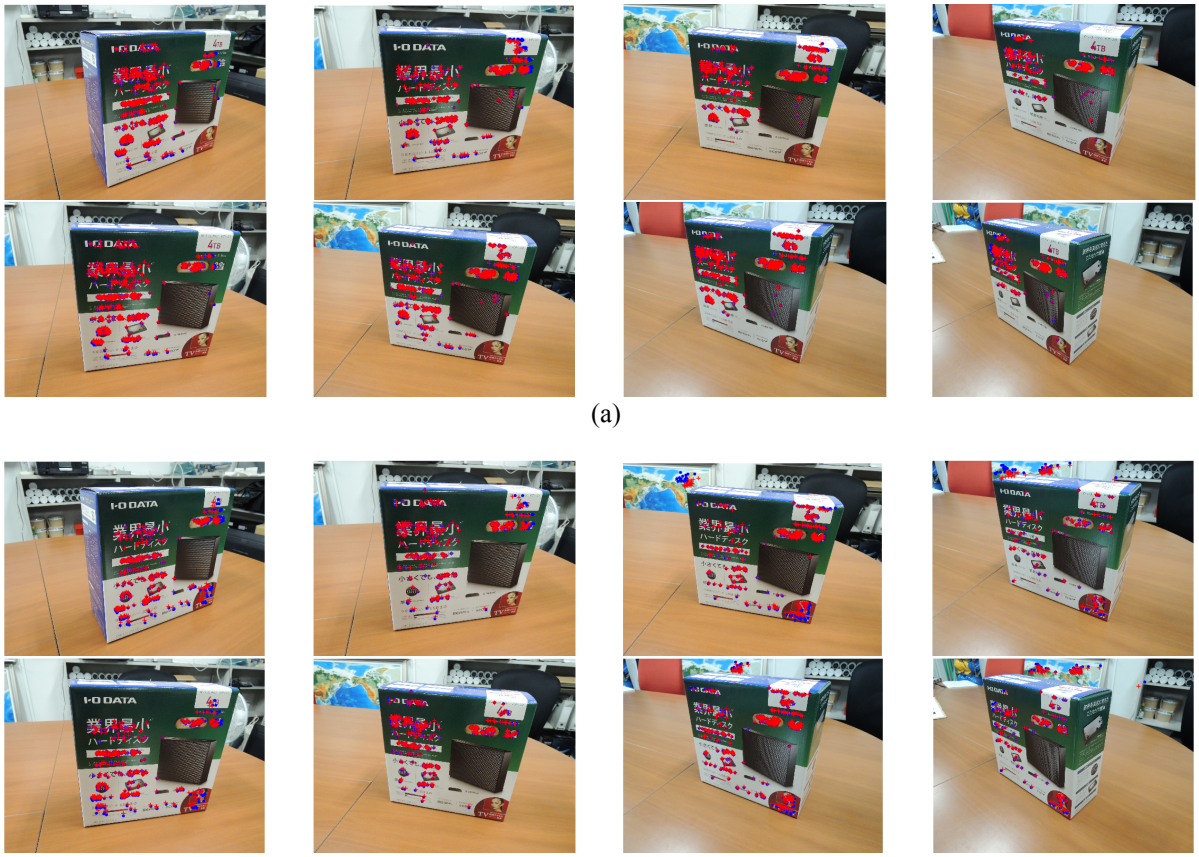
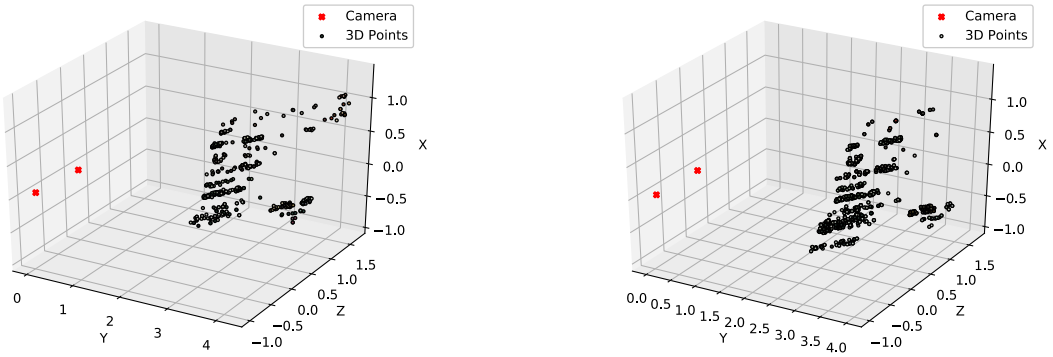


Figure 4. Reprojections of 3D points to 2D images by the reconstructed camera poses by (a) BRISK features (b) SIFT features (blue dots are measured features and red marks are projected points)

Table 2. Comparisons between BRISK and SIFT for SfM

Feature Type	BRISK (threshold t = 67)	SIFT
Time Consumed (sec)	390	398
Number of 3D Points	3093	1542
Computational Efficiency (points/sec)	7.929	3.873



(a) (b)

Figure 5. Reconstructed 3D object points and the camera poses via (a) SIFT image features (b) BRISK image features

4. CONCLUSION

3D reconstruction and camera pose recovery are important stages in optical processing. By the aid of structure-from-motion (SfM), images can be handled more efficient than before to extract the spatial information from them. In recent years, the computation cost has been one of concerns for data processing and application. The vector-based and binary-based image keypoints have been compared and discussed in the category of keypoint matching. Therefore, this paper tries to compare the computational effects of BRISK and SIFT on camera pose recovery and 3D reconstruction. The experimental results suggest that binary-based keypoint is very competitive to vector-based descriptors. Less time consumed and more 3D points produced are the main benefits according to the outcomes by applying binary-based image features for SfM. With reliable camera poses, dense matching and dense 3D reconstruction can be progressed in the future through the generation of epipolar images.

Acknowledgment

This research was supported in a part by scientific research grant-in-aid KAKENHI 19H02408.

References

Bay, H., Ess, A., Tuytelaars, T. and Gool, L.V., 2008, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110(3), pp. 346–359.
Bradski, G., Konolige, K., Rabaud, V. and Rublee, E., 2011, "ORB: An efficient alternative to SIFT or SURF," *IEEE International Conference on Computer Vision (ICCV 2011) (ICCV)*, Barcelona, pp. 2564-2571.
Calonder, M., Lepetit, V., Strecha, C. and Fua, P., 2010, "BRIEF: Binary robust independent elementary features," *In Proc. European Conference on Computer Vision; Lecture Notes in Computer Science*, Vol. 6314, pp. 778–792.
Ke, Y. and Sukthankar., R., 2004, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR,* Washington, DC, USA, pp. 27–44.
Leutenegger, S., Chli, M. and Siegwart, R. Y., 2011, "BRISK: Binary Robust invariant scalable keypoints," *2011 International Conference on Computer Vision*, Barcelona, pp. 2548-2555.
Lowe, D., 2004, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp.91-110.
Rosten, E., Porter, R. and Drummond, T., 2006, "Faster and better: A machine learning for high speed corner detection," in *Proc. European Conference on Computer Vision*, Vol. 1, pp. 430-443.