Evaluation of feature matching in aerial imagery for structure-from motion and bundle adjustment

Ke Gao, Hadi AliAkbarpour, Kannappan Palaniappan, Guna Seetharaman

Ke Gao, Hadi AliAkbarpour, Kannappan Palaniappan, Guna Seetharaman, "Evaluation of feature matching in aerial imagery for structure-from motion and bundle adjustment," Proc. SPIE 10645, Geospatial Informatics, Motion Imagery, and Network Analytics VIII, 106450J (27 April 2018); doi: 10.1117/12.2309805

SPIE.

Event: SPIE Defense + Security, 2018, Orlando, Florida, United States

Evaluation of Feature Matching in Aerial Imagery for Structure-from-Motion and Bundle Adjustment

Ke Gao[a], Hadi AliAkbarpour[a], Kannappan Palaniappan[a], and Guna Seetharaman[b]

[a]University of Missouri-Columbia, MO, USA
[b]US Naval Research Laboratory, Washington D.C., USA

ABSTRACT

Local feature matching has been proven to be successful for computer vision tasks such as Structure-from-Motion (SfM) and 3D reconstruction. Reliability of features in terms of being precisely detected and persistently matched along a sequence can have a great impact on the quality of the SfM and even on its convergence. Since many feature detectors and descriptors are exclusively designed for specific applications, it is important to find a feature detector-descriptor combination that performs well for SfM. In this paper we evaluate the quality of different image features such as FAST,[1] SIFT,[2] SURF,[3] and BRISK[4] and their effects on the Structure-from-Motion performance. To do this end, we design and perform two evaluation procedures to assess a feature matching result on a wide area motion imagery dataset. A matching result is represented in the form of feature track and a track is a collection of continuously matched feature points along the sequence. First we use the concept of Epipolar geometry to measure errors in each correspondence (matching pair). The distance from a matched feature point to the corresponding epipolar line is measured as the error metric. Second, we compute an optimized metadata from SfM using feature matching tracks and then compare it with the ground truth metadata for evaluation. Experimental results demonstrate that SURF detector combined with SURF descriptor generates the longest feature tracks while FAST detector plus SIFT descriptor produces the highest matching precision.

Keywords: feature matching, evaluation, Structure-from-Motion, 3D reconstruction

1. INTRODUCTION

Local feature matching has been proven to be successful for a wide range of computer vision tasks including Structure-from-Motion (SfM) and 3D reconstruction. This has led to a large amount of works on feature detectors and feature descriptors.

Several evaluations have been performed on the feature detectors and descriptors. Moreels et al.[5] evaluated a number of detector-descriptor combinations based on 3D objects from 144 calibrated viewpoints under three lighting conditions. They concluded that Hessian-affine detector combined with SIFT descriptor is the most robust to viewpoint changes. For illumination changes, Harris-affine detector with SIFT descriptor and Hessian-affine with shape context descriptor performed better than other combinations. Mikolajczyk et al.[6] compared the performances of a set of descriptors for the feature points under real geometric and photometric transformations. Recall and precision were introduced as the criteria for evaluation. From their observation, the performance of the descriptors are mostly independent of the feature detectors and SIFT-based descriptors perform best.

Although the SIFT-like descriptors yield competitive results, their computational costs are too high for some of the real-time applications. To resolve this issue, several binary feature point descriptors were proposed. Heinly et al.[7] analyzed the performances of three binary descriptors BRIEF,[8] ORB,[9] and BRISK[4] using SIFT and SURF as a baseline. A set of detectors were combined with these descriptors and tested on the extended version of Oxford dataset.[10] Several conclusions were made from their experiments. First, significant speed gain was

Further author information: (Send correspondence to Kannappan Palaniappan.)
Ke Gao: E-mail: kg954@mail.missouri.edu
Hadi AliAkbarpour: E-mail: aliakbarpourh@missouri.edu
Kannappan Palaniappan: E-mail: palaniappank@missouri.edu
Guna Seetharaman: E-mail: guna.seetharaman@nrl.navy.mil

Geospatial Informatics, Motion Imagery, and Network Analytics VIII, edited by Kannappan Palaniappan, Peter J. Doucette, Gunasekaran Seetharaman, Proc. of SPIE Vol. 10645, 106450J · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2309805

achieved by the binary descriptors compared to SIFT and SURF. Second, performances of the binary descriptors varied based on image transformations in the data. Third, SIFT was the best for all datasets except the non-geometric transforms. Bekele et al.[11] presented an extended evaluation of binary descriptors on Stanford Mobile Visual Search and Oxford dataset. The evaluation results demonstrate that BRISK yields higher percentage of precision and larger amount of best matches than other studied binary descriptors, while BRISK is more computationally expensive.

Many feature detectors and descriptors were exclusively designed for a specific application.[12] Their performances can be greatly affected by input data. It is most unlikely that one feature detector combined with one descriptor will outperform the others in all applications. Additionally, precise feature detection and matching are crucial for SfM and 3D reconstruction, so it is important to find a detector-descriptor combination that performs well for this application. In this paper, a set of combinations of feature detectors and feature descriptors are evaluated and compared for structure-from-motion and bundle adjustment. Section 2 describes the feature detectors and descriptors that were tested. Section 3 presents the dataset and the evaluation metrics. Section 4 summarizes the performances of the studied methods.

2. FEATURE DETECTORS AND DESCRIPTORS

Local feature is an image pattern that has distinctive characteristics from its immediate neighborhood.[13] In many applications, local features are represented by feature points. For 3D reconstruction, a reliable feature detector that can accurately localize feature points in images over time is of great importance. To create a feature descriptor, a local region around a feature point is extracted and converted into a high-dimensional vector, which acts as a feature descriptor. An robust descriptor is invariant to certain image transformations, such as translation, rotation, scale, etc. Then a feature descriptor in the sensed image are matched against the descriptors in the reference image to produce correspondence between feature points.

2.1 Feature Detectors

2.1.1 FAST

Features from Accelerated Segment Test (FAST)[1] is a corner detection algorithm. For a candidate point p that is to be identified, let I_p denotes its intensity. FAST corner detector considers a circle of 16 pixels around p to determine whether it is a corner. The candidate point p is classified as a corner if the intensity of a set of N contiguous pixels in the circle around it are all larger than I_p plus a threshold t or all smaller than I_p minus a threshold t. To operate the high-speed test that reduces the number of non-corner points, only four of the selected pixels in the circle are examined. If three out of four pixels are brighter than $I_p + t$ or darker than $I_p - t$, the full circle (16 pixels) examination can be applied. Despite the high performance of this detection algorithm, several weaknesses still exist. First, not as many non-corner points can be excluded when N is smaller than 12. Second, the efficiency of corner detection is affected by the choice and ordering of pixels as well as the distribution of corner appearances. Third, results of high-speed tests are thrown away. To address these problems, a machine learning approach is introduced. Another problem is that multiple feature points are adjacent to each other. It can be solved by applying the non-maximal suppression. The major advantage of FAST corner detector is its high-speed performance. It is faster than many existing corner detection algorithms.

2.1.2 SIFT

Scale-Invariant Feature Transform (SIFT)[2] uses Difference of Gaussians (DoG) for keypoints (points of interest) detection in different scales. The image is convolved with Gaussian filters of various σ values. Difference of Gaussian is then obtained by taking the difference between two successive Gaussian-blurred images. Then the DoG at different scales are searched for local maxima/minima to identify the candidate keypoints. To obtain a more accurate keypoint localization, the candidate keypoints are refined by calculating the interpolated location of maxima/minima, eliminating edges and low-contrast keypoints. After the keypoints are located, an orientation is assigned to each of them so rotation invariance can be achieved in the matching process. A neighboring region around the keypoint is taken and gradient magnitude and orientation of every pixels in the region are computed to form an orientation histogram. The peak in the histogram correspond to the dominant orientation. If there exists multiple peaks in the histogram, additional keypoints are created at the same location and scale but different orientations.

2.1.3 SURF

Speeded Up Robust Features (SURF)[3] feature detector is based on SIFT algorithm but increase the computation efficiency. SURF uses a box filter to approximate Laplacian of Gaussian. It is faster because filtering an image with a square-shaped filter can be done with the help of the integral images. SURF uses the Hessian matrix whose determinant determines the location and scale of feature points.

2.1.4 BRISK

Binary Robust Invariant Scalable Keypoints (BRISK)[4] is a corner detection algorithm inspired by AGAST,[14] which is an extension of FAST[1] by increasing the speed. To achieve the scale invariance, BRISK uses the scale-space pyramid consisting of several octaves and intra-octaves. The octaves are generated by progressively down-sampling the original image. The scale-space interest point detection starts with applying FAST 9-16 detector on each octave and intra-octave separately. Then non-maxima suppression is performed not only in the layer that the potential region of interest lies, but the immediately-neighboring layers above and below. As the final step, the location of the keypoint is re-interpolated across different scales.

2.2 Feature Descriptors

2.2.1 SIFT

A 16 x 16 region around each feature point is taken and each region is divided into 16 subregions with 4 x 4 size. An 8-bin orientation histogram of each subregion is generated using gradient magnitude and orientation of samples in the subregion. So the descriptor for a feature point is created by combining all of the orientation histograms from 16 subregions. The descriptor is normalized to unit length so that it is more robust against illumination changes.

2.2.2 SURF

Orientation is estimated and assigned to each feature point to achieve rotation invariance. Wavelet responses are computed in x- and y-directions inside a circular neighborhood of a point of interest. Gaussian weights are applied to the obtained responses. The weighted responses are plotted in a two-dimensional space. Then a sliding orientation window is created and the sum of all responses in it is used for estimating the dominant orientation. To provide description of a feature point, a square region around the point is extracted and transformed along the dominant orientation. The square region is divided into 4x4 sub-regions. Wavelet responses are taken in each sub-region and formed into a vector as the feature descriptor.

2.2.3 BRISK

BRISK uses a binary string as the feature descriptor. It is generated from the result of multiple brightness comparison tests. To sample the local region around a feature point, a deterministic pattern that consists of N points is utilized. The sample points form concentric circles centered at a feature point. The distance between each sample point to the corresponding feature point determines the standard deviation σ of a Gaussian filter. Each Gaussian filter is applied to the local area to eliminate the aliasing effects. To achieve scale and rotation invariance, each sampling pattern is scaled and rotated by its characteristic pattern orientation. The orientation is estimated from a set of long-distance pair comparisons. For each pair, its local gradient is determined by computing the intensity difference between two sample points, weighted by their displacement in position. Then the average gradient of all the long-distance pairs is the pattern orientation.

3. EVALUATION

3.1 Dataset

The Wide Area Motion Imagery (WAMI) dataset were collected (by Transparent Sky) using an aircraft with on-board pose sensors flying over five different urban areas including down- town Albuquerque, NM. Two sample images from the Albuquerque Dataset are shown in Fig 1.

(a) (b)

Figure 1. Sample Images from Albuquerque Dataset.

3.2 Evaluation

To evaluate the quality of various feature detector-descriptor combinations, we design and perform two evaluation metrics (Fig 2). First, we use the concept of Epipolar geometry to measure errors in each correspondence (match pair) in the dataset (Fig 2(a)). Second, we compute an optimized metadata from SfM using feature matching tracks and then compare it with the ground truth metadata using the fundamental matrix (Fig 2(b)). The feature extraction methods were run on the WAMI dataset. After feature extraction, each image frame was compared to its next frame within the sequence in order to find matches. In each match, the epipolar line corresponding the first frame was computed (using available groundtruth camera parameters) and plotted on the second frame. The distance between the epipolar line and the feature point on the second image was measured as an error metric. A similar error was obtained by drawing the epipolar line of the second frame over the first one and measuring the distance. The mean of these two errors were assigned as the error corresponding to that match (pair of features).

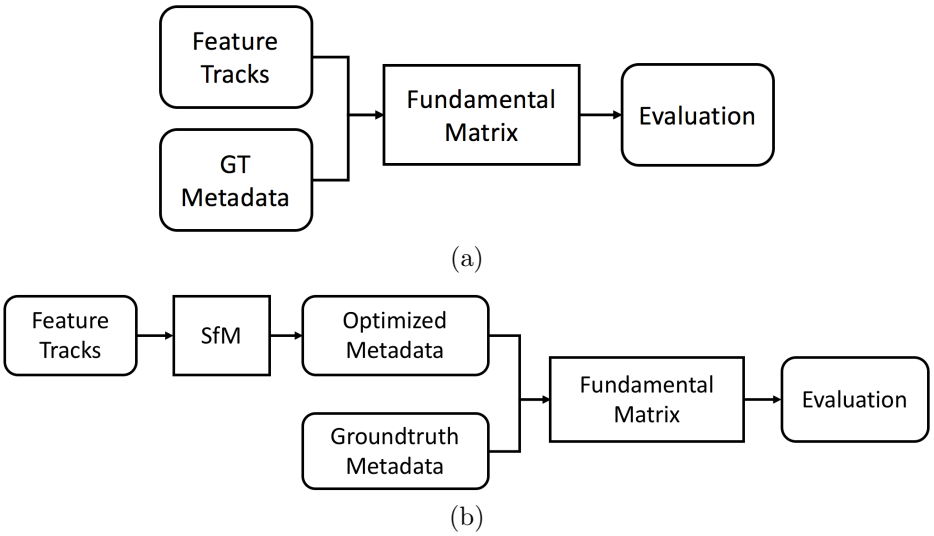
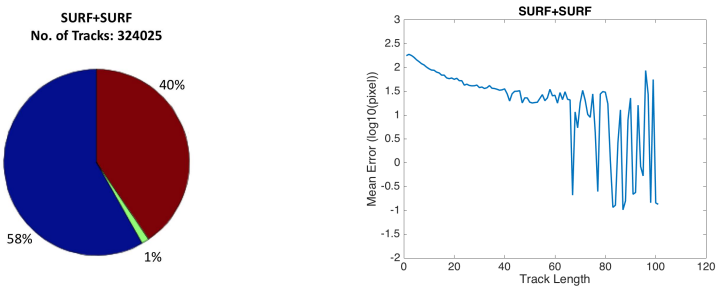


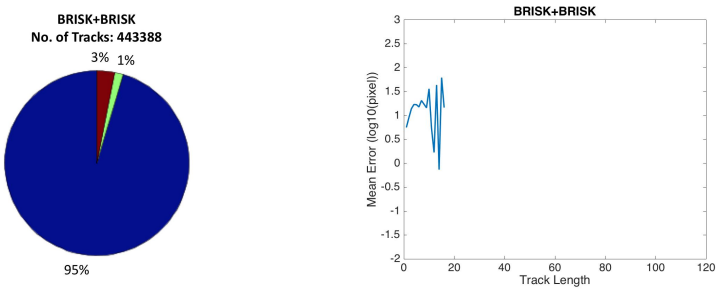
Figure 2. Evaluation metrics

Left column in Fig 3 depicts the percentage of errors for different combination of feature extraction and feature descriptor methods. In addition to these overall error percentages, we have also evaluated errors with respect to the tracks lengths. A track is basically a collection of continuously matched features along the sequence. In

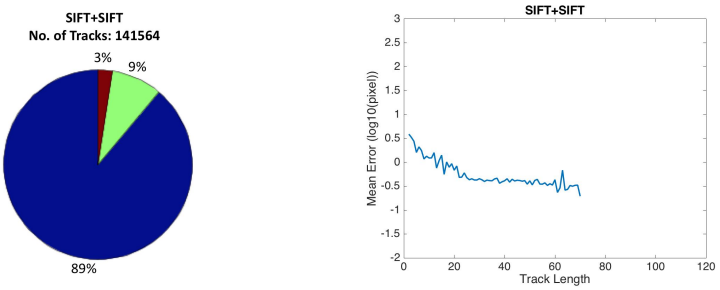
some algorithms such as SfM and BA, having long tracks with low errors is very crucial for converging to global solution. Therefore, we have measured the error in pixel for each track by computing the average error of all matches in that track. Right column in Fig 3 shows the changes in error values (base 10 logarithm) with respect to the track length.



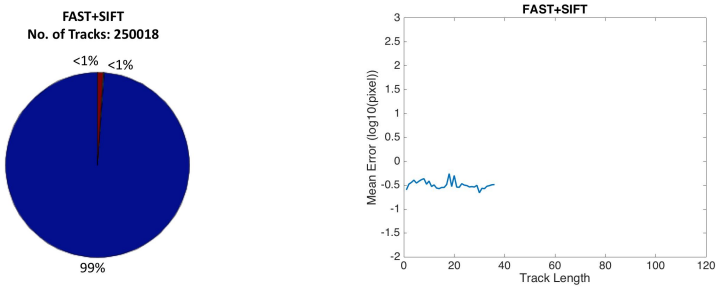
SURF feature detector + SURF feature descriptor



BRISK feature detector + BRISK feature descriptor



SIFT feature detector + SIFT feature descriptor



FAST feature detector + SIFT feature descriptor

Figure 3. Error evaluation. Left: Percentage of match error in pixel (blue: error ≤ 1; green: $1 <$ error ≤ 2; red: error ≥ 3). Right: Errors of the base 10 logarithm with respect to track lengths.

As shown in Fig 3, SURF detector with SURF descriptor generates the longest tracks. The maximum length

reaches around 100 frames which is almost half the length of the Albuquerque dataset sequence (215 frames in total). However, its feature matching precision is the lowest among four testing results. 40% of the matched features have an error of more than 3 pixels. BRISK feature detector and descriptor generate a high matching precision with 95% of features under 1 pixel error. Its track length is the shortest. The longest track is only around 16 frames. SIFT feature detector and descriptor shows promising result in terms of both precision and track length. The maximum track length is around 70 and only 12% of feature points have an error larger than 1 pixel. FAST detector combined with SIFT descriptor shows the highest matching precision. 99% of features have an error smaller or equal to 1 pixel. The longest track has only half the length compared to the SIFT+SIFT combination.

4. CONCLUSION AND FUTURE WORK

In this paper, a set of feature matching results are evaluated and compared in aerial imagery for SfM and bundle adjustment. Feature matching results are in the form of feature tracks. In our proposed approach, feature points are extracted from each frame in a sequential image dataset. For the feature points in a frame, a descriptor is created and compared to the features in the next frame. The successive matches merged together become feature tracks. The precision of a detector-descriptor combination is reflected by the errors of feature tracks they generated compared to the epipolar lines in each frame. The experimental results illustrate that SURF detector combined with SURF descriptor generates the longest feature tracks but the matching precision is low. FAST detector combined with SIFT descriptor produces the highest precision while the lengths of its tracks are shorter. For the future work, we will include more state-of-the-art feature detectors as well as descriptors for matching evaluations.

REFERENCES

[1] Rosten, E. and Drummond, T., "Machine learning for high-speed corner detection," in [*European conference on computer vision*], 430–443, Springer (2006).

[2] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *International journal of computer vision* **60**(2), 91–110 (2004).

[3] Bay, H., Tuytelaars, T., and Van Gool, L., "Surf: Speeded up robust features," in [*European conference on computer vision*], 404–417, Springer (2006).

[4] Leutenegger, S., Chli, M., and Siegwart, R. Y., "Brisk: Binary robust invariant scalable keypoints," in [*2011 International conference on computer vision*], 2548–2555, IEEE (2011).

[5] Moreels, P. and Perona, P., "Evaluation of features detectors and descriptors based on 3d objects," *International Journal of Computer Vision* **73**(3), 263–284 (2007).

[6] Mikolajczyk, K. and Schmid, C., "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence* **27**(10), 1615–1630 (2005).

[7] Heinly, J., Dunn, E., and Frahm, J.-M., "Comparative evaluation of binary features," in [*Computer Vision–ECCV 2012*], 759–773, Springer (2012).

[8] Calonder, M., Lepetit, V., Strecha, C., and Fua, P., "Brief: Binary robust independent elementary features," in [*European conference on computer vision*], 778–792, Springer (2010).

[9] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "Orb: An efficient alternative to sift or surf," in [*Computer Vision (ICCV), 2011 IEEE International Conference on*], 2564–2571, IEEE (2011).

[10] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Van Gool, L., "A comparison of affine region detectors," *International journal of computer vision* **65**(1-2), 43–72 (2005).

[11] Bekele, D., Teutsch, M., and Schuchert, T., "Evaluation of binary keypoint descriptors," in [*Image Processing (ICIP), 2013 20th IEEE International Conference on*], 3652–3656, IEEE (2013).

[12] Guo, Y., Bennamoun, M., Sohel, F., Lu, M., Wan, J., and Kwok, N. M., "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision* **116**(1), 66–89 (2016).

[13] Tuytelaars, T., Mikolajczyk, K., et al., "Local invariant feature detectors: a survey," *Foundations and trends® in computer graphics and vision* **3**(3), 177–280 (2008).

[14] Mair, E., Hager, G. D., Burschka, D., Suppa, M., and Hirzinger, G., "Adaptive and generic corner detection based on the accelerated segment test," in [*European conference on Computer vision*], 183–196, Springer (2010).