ABSTRACT
In this paper an evaluation of state-of-the-art binary keypoint descriptors, namely BRIEF, ORB, BRISK and FREAK, is presented. In contrast to previous evaluations we used the Stanford Mobile Visual Search (SMVS) data set because binary descriptors are mainly used in mobile applications. This large data set does provide a lot of characteristic transformations for mobile devices, but no ground truth data. The often used Oxford data set is used only for validation purposes. We use ratio-test and RANSAC (RANdom SAmple Consensus) for evaluation and present results for accuracy, precision and average number of best matches as performance metrics. The validity of the results is also checked by evaluating these binary keypoint descriptors on Oxford data set. The obtained results show that BRISK is the keypoint descriptor which gives highest percentage of precision and largest number of best matches among all the binary descriptors. Next to BRISK is FREAK, which offers comparably good result.

Index Terms—binary descriptors, matching, recognition, invariance, evaluation, mobile feature tracking

1. INTRODUCTION
The increasing number of mobile applications which are based on image registration and recognition, e.g., Augmented Reality or comparison shopping, led to a significant amount of publications of novel image registration techniques. A common approach for image registration is to compute significant points, so called keypoints or feature/interest points in the image, and a description of the neighborhood of each keypoint, which is invariant to certain transformations. These descriptors are then compared to descriptors extracted from database images in order to find a matching image in the database. The existence of accurate and reliable keypoint detectors together with appropriate descriptors is therefore an important step. This paper focuses on the evaluation of feature descriptors for mobile image registration and recognition applications.

One of the most well-known keypoint descriptors is SIFT (Scale Invariant Feature Transform) [1] which detects keypoints based on Difference of Gaussians (DoG). Although SIFT has been published several years ago, it still yields competitive results to state-of-the-art techniques. Apart from SIFT, several SIFT-like descriptors have been published, which involve more or less modifications, e.g., ASIFT [2] or PCA-SIFT [3]. SURF (Speeded-Up Robust Features) [4] is one of the most popular modifications, which yields similar matching performance, but faster computation. However, processing time for SIFT-like descriptors is still too high for real-time applications on mobile devices with limited computing power and memory capacity.

Binary keypoint descriptors aim to filling this gap. They show similar performance to SIFT-like descriptors, while having significant lower computational costs. The idea behind binary descriptors is that each bit in the descriptor is independent and the Hamming distance can be used as similarity measure instead of, e.g., the Euclidean distance. The four most recent and promising binary feature descriptors are (1) Binary Robust Independent Elementary Feature (BRIEF) [5], (2) Oriented Fast and Rotated BRIEF (ORB) [6], (3) Binary Robust Invariant Scalable Keypoints (BRISK) [7] and (4) Fast Retina Keypoint (FREAK) [8].

Most publications concerning feature descriptors refer to Mikolajczyk and Schmid [9], who compared the performance of descriptors computed for feature points on the Oxford data set. Mikolajczyk and Schmid introduced recall and precision as performance measures and compared different types of descriptors, but no binary ones. They conclude, that the ranking of descriptors is mostly independent of the feature point detector and that SIFT-like descriptors yield best performance. More recently Heinly et al. [10] presented an evaluation of the binary feature descriptors BRIEF, ORB and BRISK as well as SURF and SIFT. Their evaluation is based on an extended version of the Oxford data set and they also analyzed the performance of different detector and descriptor combinations. Their main conclusions are that (1) descriptors should be adapted to the transformations present in the data, (2) both detector and descriptor should be invariant to the same set of transforms and (3) speed gains achieved by binary descriptors result (at worse) in marginal matching performance penalties. They also introduced additional evaluation measures compared to Mikolajczyk and Schmid, i.e., an entropy measure and a measure for the frequency of candidate measures. However, these measures do not seem to effect the overall results. Chandrasekhar et al. [11] presented the Stan-
ford Mobile Visual Search data set. This data set overcomes some limitations of other data sets, such as the Oxford data set, e.g., the lack of realistic ground truth reference data or the use of camera phones for query images.

We extend the evaluation of Heinly et al., so that we (1) evaluate an additional binary descriptor, namely FREAK, and (2) use the database proposed by Chandrasekhar et al., which is from our point of view more appropriate for evaluating binary descriptors.

2. DATA SETS

The evaluation has been performed on two data sets. The Oxford data set proposed by [9] and the Stanford Mobile Visual Search (SMVS) [11] dataset. The Oxford data set contains eight different scenes where each scene includes one of the following changes: brightness, viewpoint, rotation and scale, JPEG compression or blur. The data sets offer ground truth to each scene, which makes it possible to estimate true errors. However, real life scenes are much more complex. The SMVS data set includes a wide range of images (3300 query images for 1200 classes across 8 image categories). The images are captured with several different camera phones including some digital cameras and are captured indoors and outdoors under widely varying lighting conditions over several days including foreground and background clutter. The drawback of the SMVS data set is that the data set has been generated for mobile search, i.e. ground truth is a reference image, but no transformation ground truth is given. In order to compare descriptors without transformation ground truth, we use different tests which are explained in Section 3.

In our experiments we focus on the SMVS data set and use for comparison the Bark subset from the Oxford data set which includes scale changes resulting from varying camera zoom.

3. EVALUATION

Our evaluation is performed on four binary descriptors namely: BRIEF, ORB, BRISK and FREAK. In addition SIFT is used as reference descriptor. SURF detector is used for BRIEF, BRISK and FREAK descriptors. ORB detector is used for ORB descriptor, because compared to the other descriptors, ORB descriptor needs keypoint orientation information. Also, the evaluation of Heinly et al. [10] showed that the ORB/ORB pairing outperformed the SURF/ORB pairing in almost all cases. On the other hand, SURF keypoints are invariant to rotation and scale changes which makes them suitable to meet the requirements for BRISK and FREAK descriptors which are also scale and rotation invariant. The authors of BRIEF [5] also compute their descriptors on SURF keypoints confirming our choice of keypoint detector.

The first step in the evaluation is detecting keypoints in both the reference and query images. In order to compare descriptor performance independent of the number of keypoints, we limit the number to at maximum 500. Then the descriptors using the detector-descriptor combinations discussed above are computed. Feature matching is done using a brute-force matcher. Two best matching points for each keypoint are selected based on the distance between their descriptors. If this measured distance is very low for the best match, and much larger for the second best match, the first match is selected as a good one since it is unambiguously the best choice. Reciprocally, if the two best matches are relatively close in distance, the probability for a mistake is high. In this case, both matches are rejected. This test is called Ratio Test and verifies that the ratio of the distance of the best match over the distance of the second best match is not greater than a given threshold. Lowe [1] shows the probability density functions for correct and incorrect matches in terms of the ratio test. A threshold value of 0.8 eliminates 90 % of the false matches while discarding less than 5% of the correct matches.

The next step in the evaluation is the RANdom SAmple Consensus (RANSAC). When two cameras observe the same scene, they see the same elements but under different viewpoints. This method will allow exploiting the epipolar constraint between two views to match image features more reliably. The principle is simple: when matching keypoints between two images, accept only those matches that fall onto the corresponding epipolar lines. However, to be able to check this condition, the fundamental matrix must be known, or correct matches are needed to estimate this matrix. The SMVS data set does not provide transformation ground truth, therefore we jointly compute fundamental matrix and a set of good matches. These good matches are then assumed to be correct matches.

The final step in the evaluation is to calculate precision and recall in percentage. The ratio of the number of correct matches to the symmetrical matches gives the precision of matching. Recall of matching is given by the ratio of the number of correct matches to the number of matches which have passed the ratio test.

4. RESULTS AND DISCUSSION

The evaluation method using Stanford Mobile Visual Search (SMVS) data set uses four performance evaluation parameters namely: average number of keypoints, precision in percentage, recall in percentage and the average number of best matches. In order to equalize the number of keypoints, the parameters of the detectors involved are tuned in a way that approximately 500 correspondences are detected in the reference image. If the number of detected correspondences is larger than 500, the first 500 keypoints are taken. This allows to have a fairly equal number of feature points for each algorithms to be compared. Figure 1 shows the average number of points for the binary descriptors with respect to the various sub data sets of SMVS. As seen in the bar graph, ORB and
SIFT yield in general between 400 and 500 feature points, whereas the other descriptors do compute significantly less feature points. The following tests are used to compute correct matches according to Sec. 3.

Ratio test is the first evaluation method. The ratio of the distance of the best match to the second best match will be taken and if this ratio is greater than or equal to 0.9 both of the matches will be ignored, if not the best match will be kept. This helps to eliminate considerable amount of false matches.

The next test is RANSAC. The idea is to find the best projective relationship between the reference and the query keypoints by computing a fundamental matrix from eight random sample points. The output of this test is the set of final best matches or correct matches. Then precision is calculated by taking the ratio of the number of best matches to the number of matches that passes the ratio test. Recall is calculated by taking the ratio of the number of best matches to the number of correspondences (number of matches before the ratio test).

Figure 2 below shows the precision in percentage for the basic descriptors with respect to the sub data sets of the SMVS data set.

As seen in Fig. 2, BRISK and FREAK give best precision values in percentage while BRIEF gives least precision percentage value.

Yet another important measure of comparison is the average number of best matches which is calculated by multiplying the average number of keypoints with the percentage of accuracy. Accuracy is the ratio of the number of best matches to the number of keypoints that are matched which includes the matches eliminated by ratio test. Figure 3 shows the recall in percentage for the SMVS sub data sets. BRISK and FREAK offer the highest percentage of recall for most data sets except for Landmarks and Print which is originated from the nature of the images in the data sets. Fig. 4 shows the average number of best matches.

As can be inferred from the graph and leaving SIFT as a reference, BRISK is performing best except for sub data sets Business cards, Landmarks and Print. On the other hand BRIEF gives the least number of best matches. Next to BRISK is FREAK which offers the largest number of best matches on Business cards and Print sub data sets, while performance of ORB and BRIEF is significantly lower except for the sub data set Landmarks.

Lowe [1] stated that three keypoints may be sufficient for reliable recognition of objects. So, the results of the descriptors which are less than 20 in Fig. 4 are still important and cannot generally be considered as poor performance. But algorithms with higher number of best matches are more robust for an application at hand.

The results in Print and Landmarks sub data sets give considerably low recall and precision as well as a low average number of best matches. This is due to the nature of the images as they contain highly distorted query images. Some examples of images are shown in Fig. 5.

In another evaluation the four binary descriptors (BRIEF,
ORB, BRISK and FREAK) together with the state-of-the-art detector/descriptor SIFT have been evaluated on Oxford data set provided and described by Mikolajczyk et al. [9]. Specifically, Bark sub data set is selected because of its similarity with SMVS data set including scale changed query images. The results for the number of best matches for different scale change values are shown in Fig. 6. It shows that BRISK offers the highest number of best matches next to SIFT which agrees well with the results from SMVS data set. This confirms the validity of the results on different data sets which leads to the same rank of binary descriptors BRISK being the first while FREAK, ORB and BRIEF are ranked 2nd, 3rd and 4th respectively.

5. CONCLUSION

In this paper we have presented an experimental evaluation of binary keypoint descriptors on Stanford Mobile Visual Search and Oxford data set. The desirable goal was to compare and rank those descriptors based on their performance and compare them with the state-of-the-art descriptor SIFT. Note that the evaluation was designed for matching and object recognition in mobile applications.

The comparison presented here aimed to give an impression on evaluations which are based on real world scenarios. In other words, query images without ground truth matrices which offer the transformation between the reference image and the query image. Apart from this, SMVS data set also offers real image distortions on a large number of images. To test the algorithm at hand typically each sub data set contains 500 query images.

From evaluation results obtained BRISK is recommended as the best binary keypoint descriptor yielding the largest amount of best matches which is closely comparable with the results of SIFT. It should be noted, that BRISK needs significantly more computation effort compared to BRIEF and ORB [10]. Also FREAK descriptors are faster to compute compared to BRISK, while simultaneously needing less memory load.

We conclude, that BRISK yields best results on the SMVS data set, while FREAK offers only slightly less performance but faster computation.

6. REFERENCES


