Evaluation of Salient Point Methods

Song Wu Michael S. Lew
LIACS Media Lab, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
{wusong, mlew}@liacs.nl

ABSTRACT
Processing visual content in images and videos is a challenging task associated with the development of modern computer vision. Because salient point approaches can represent distinctive and affine invariant points in images, many approaches have been proposed over the past decade. 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 salient point detectors and descriptors. We begin by comparing diverse salient point algorithms (SIFT, SURF, BRIEF, ORB, FREAK, BRISK, STAR, GFTT and FAST) with regard to repeatability, recall and precision and then move to accuracy and stability in real-time video tracking.

Categories and Subject Descriptors
Computing methodologies, computer vision

General Terms
Algorithms, Performance, Experimentation.

Keywords
Salient point methods, evaluation, video tracking

1. INTRODUCTION
Salient point detection and feature descriptors have become prevalent in diverse areas in computer vision and multimedia information retrieval [7,8,14,15]. A well known algorithm is the Scale Invariant Feature Transform (SIFT) by Lowe [1] which has been shown repeatedly to have good accuracy from the research literature. However, it has also been noted that the generated 128 dimensional descriptor may lead to relatively large descriptors and lower computational efficiency. A faster approach named SURF (Speeded Up Robust Feature) was developed by Bay et al [2].

In contrast to the real value descriptors (i.e. SIFT, SURF), binary string descriptors were developed with the aim to compute the feature descriptors more efficiently. The representative method of binary string descriptors is BRIEF [3], which is constructed by intensity comparison of pixel-pairs. Since BRIEF is not invariant to scale and orientation, a series of modified descriptors had been developed. Rublee et al [4] designed ORB (oriented FAST and Rotated BRIEF) as an alternative to SIFT and SURF. The ORB detector is an extension of FAST and the ORB descriptor is an improvement of the BRIEF descriptor. The drawback regarding scale invariance still exists. The BRISK (Binary Robust Invariant Scalable Keypoints) approach [5], which is a method for keypoint detection and description, performs much faster than the more well-established SIFT and SURF approaches. A recent new keypoint descriptor named FREAK (Fast Retina Keypoint) [6] was designed to enhance the performance of existing keypoints descriptor. It was formed by comparing image intensities over a retinal sampling pattern.

The focus of this paper is to assess a wide set of older (i.e. SIFT) and more recent salient point approaches. The main contributions of this work are insights based on the precision-recall graphs from the main image transformations and also the comparison of the different approaches in video tracking. It should be noted that regarding objectivity, the authors did not submit any of their own algorithms for this evaluation.

1.1 Related Work
Schmid et al [7] used the measure of "repeatability rate" and "information content" to show the performance of different detectors. In addition, several works have been done on the evaluation of local descriptors by measuring the accuracy of matching and recognition [8]. Recall and precision are two widely used indicators to denote the performance under various affine translations. Accuracy and speed trade-offs [9] have been studied where different indexing structures were employed (such as approximate kd trees). Gauglitz et al. in [10] presented a comparison of different salient approaches on object tracking in video sequences. As far as we know, this is the only work which covers the newer descriptors such as BRIEF and FREAK regarding precision-recall and stability in the form of trajectory jitter noise.

2. DETECTORS AND DESCRIPTORS
SIFT and SURF are two of the most influential methods. However recent new binary string features have received significant attention. These are generated by comparing the intensity of pixel-pairs and use the Hamming distance (bitwise XOR followed by a bit count) for matching instead of Euclidean distance.

SIFT [1]: The implementation of SIFT begins by building the scale space which approximates the Laplacian-of-Gaussian function by the computationally efficient Difference-of-Gaussian function. It searches extrema over all scales and then eliminates the potential points which are sensitive to edge response. The orientation is assigned to each stable point according to the local image gradient direction. Furthermore, it accumulates the orientations of a 16x16 neighborhood sample points around the keypoint location into orientation histograms by summarizing the contents over 4x4 sub-regions. A 128 dimensional descriptor vector is finally generated for each feature point.

SURF [2]: The box-filters together with integral images are exploited to approximate the Hessian matrix which is used to

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502179

measure the interest points. The scale space of SURF is established by up-scaling the size of the box-filter. The final salient points are assigned orientations which are calculated by summing the Haar-wavelet responses in a circular neighborhood of the salient point. In addition, the SURF descriptor is created by merging the values and absolute values of Haar-wavelet responses over a square region centered on and oriented along the salient point.

FAST detector [11]: A sample scheme of FAST corner detection is based on a circle (three pixels radius) of sixteen pixels around the candidate point, if a set of twelve contiguous pixels in the circle which are all brighter or all darker than the intensity of the point pixel value plus a threshold, the point will be classified as a corner point. Furthermore, a decision tree is generated by machine learning on training sets, and the rules of the FAST corner point classification are established to speed up the procedure.

STAR detector [12]: The STAR detector algorithm uses a simplified center-surround filter at all locations and all scales, and it detects the extrema in a local neighborhood. For each obtained extrema (based on the non-maximum suppression method, which is the same as SURF) the location of the potential points can be fixed. Furthermore, through computing the Harris measure for the potential points, those with strong edge response will be eliminated.

GFTT detector [13]: GFTT derived from an image motion model, is used as a method for feature selection, tracking and monitoring, and it performs well under image affine transformation. According to the proposed feature selection criteria, a candidate point is accepted if it is defined as a good feature which can be tracked well.

BRIEF descriptor [3]: BRIEF is a binary string feature descriptor. With regard to the located patches in an image, Gaussian smoothing is first introduced to reduce the effect of noise sensitivity so that it can achieve good performance in complex scenes. A small number of pixel-pairs are pre-selected randomly from a Gaussian distribution around the smoothed patch center and the BRIEF binary string descriptor is produced via the intensity comparison of pixel-pairs.

ORB [4]: ORB applies the FAST detector to find potential salient point locations on an image pyramid. It orders the detected points according to the Harris corner measure, and picks the top set of points as salient points. The direction of points is computed using intensity centroid. The ORB descriptor improves the BRIEF descriptor and compares the intensity of patch-pair to form the binary string vector. Then, the combination of learning and greedy search is further introduced to reduce correlation and minimize variance in the binary tests.

BRISK [5]: In the implementation of BRISK, salient points are detected using FAST within layers of the image pyramid as well as in continuous layers. BRISK presents a novel sampling pattern which consists of sample points equally distributed on concentric circles centered around the saline point, and it determines the orientation by computing local intensity gradients as well as generating a binary descriptor by comparing pairwise intensities.

FREAK descriptor [6]: The FREAK descriptor is inspired by the human retina system. FREAK samples pairs of pixels over a retinal sampling pattern and then compares their intensities. The direction is established by summing the local intensity gradients over selected pairs with symmetric receptive fields to the center of sampling pattern.

3. EXPERIMENTS

Our experiments use international public test sets in the domains of both image matching and video tracking and several evaluation measures from the research literature. The software is downloadable at http://press.liacs.nl/researchdownloads/. The experiment environment for the evaluation: AMD 64*2 Dual Core Processor (2.41GHz), and 3.2GB of RAM.

3.1 Detector Evaluation

The evaluation of different salient point detectors used the dataset provided by Mikolajczyk and Schmid [8, http://www.robots.ox.ac.uk/~vgg/research/affine/], which contains eight groups image samples with various transformations (rotation, viewpoint, scale, JPEG compression, illumination and image blur). Each group is consist of six texture or structured scene images, as well as the ground truth homography between the reference image and the transformed image.

One important evaluation measure from the research literature is repeatability [7, 8]. The repeatability score is calculated as the ratio between the number of correspondences and the minimum total number of m_1 and m_2 where m_1, m_2 denotes the number of points in reference and query images after projecting reference image points by homography and removing points outside common area.

$$\text{Repeatability}=C(m_1, m_2)/MIN(m_1, m_2)$$

$C(m_1,m_2)$ is the number of correspondences between m_1 and m_2. Overlap error is used to identify the correspondence. For a keypoint region in query image which is the nearest one to a projection keypoint region by using homograph in reference image, if the ratio between the intersection of two regions and the union of the two regions is larger than overlap error, it will be considered a correspondence. We compute the average repeatability scores on the whole dataset, thus, the detection performance of each method can be estimated in a comprehensive perspective. The trend of average repeatability under variant overlap error (in the range from 0.5 to 0.9) is shown in Figure 1.

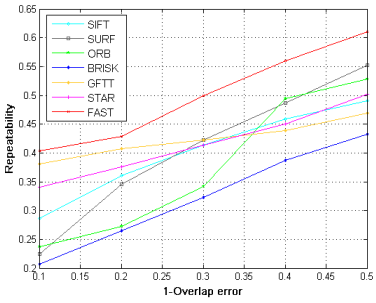

Figure 1. Comparison of various detectors using repeatability

Figure 1 illustrates that an increase in the repeatability scores is clearly indicated when overlap error becoming smaller. We also can notice that the FAST detector had the highest repeatability and the BRISK detector obtained the lowest score. All detectors can reach a stable and well performance when the value of overlap error is 0.5, the overlap error will be set at 0.5 to identify the correspondence in the following experiments.

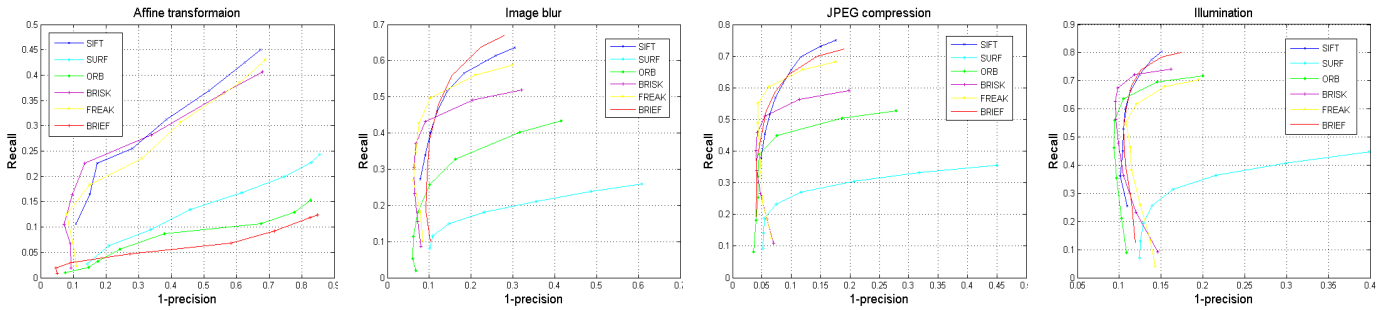

Figure 2. Comparison of various descriptors using recall vs 1-precision under different image degradation

Since different salient point detection mechanisms result in different time complexity, and different quantity of feature points can be extracted from the same image, time comparison should be compared statistically. We applied different types of detectors to various test images, in order to determine statistically significant results. The number of average detected points and the time cost of compared salient point methods are statistic in Table 1.

From the demonstration of Table 1, it reveals that the most efficient detector is FAST. FAST detected the largest number of salient points which is ten times than that obtained by other detectors. Moreover, the time cost is only 3.2(s) for the total 502924 points. The most time-consuming detector is SIFT which spends almost average 9(ms) on per point detection. The average time cost of SURF, BRISK, GFTT, and STAR is almost in the same order all with standard deviations less than 0.01.

Table 1. Comparison of average detection time

Method	Detection time(s)	Number of points	Per point time cost(ms)
SIFT	44.958	51788	8.68
SURF	62.929	148470	0.424
ORB	10.472	17336	0.604
BRISK	16.543	47236	0.35
GFTT	12.497	35000	0.357
STAR	9.69	29426	0.329
FAST	3.23	502924	0.0064

3.2 Descriptor Evaluation

The same dataset mentioned in the repeatability experiments is utilized in this part. Note that some of the salient point detectors from the previous section do not define descriptors and are not compared here. The evaluation starts by extracting salient point features from reference image and establishing KD-tree, or LSH index for them. Then, extract features from the query image and match them against the features of each reference image based on the approximate nearest neighbor method. In the procedure of matching descriptors in one pair-image, KD-tree index is established for SIFT, SURF descriptors and use Euclidean distance to realize the matching of real value descriptors.

In order to increase the matching accuracy, the Nearest Neighbor Distance Ratio (NNDR) is used as the matching strategy to find the similar descriptors in the image pairs. NNDR defines that two points will be considered a match if $||D_A-D_B||/||D_A-D_C||<threshold$, where D_B is the first and D_C is the second nearest neighbor to D_A. In addition we also computed recall and the 1-precision [8].

The noise is introduced by image degradation which results in the increase of distance between similar descriptors. Thus, we varied the value of *threshold* in the NNDR from 0.1 to 0.9 to obtain the curves about the tendency of average result of recall and 1-precision under each transformation, as shown in Figure 2. We can see that recall increases for an increasing threshold of NNDR. All descriptors perform better on image changes (blur, JPEG compression and illumination) than affine deformation. However, SIFT, BRISK and FREAK show good performance for all image degradations. Toward to the affine transformation, the curves obtained by different detectors could be classified as two groups: one is consist of SIFT, BRISK and FREAK and the other one contains SURF, ORB and BRIEF. The trend of the curves indicates that the descriptors created by SIFT, BRISK and FREAK are more robust and distinctive than SURF, ORB and BRIEF. This is mainly because that BRIEF descriptor conducted only by pixel-pair intensity comparison is not affine invariant, while ORB descriptor as an improved BRIEF is rotation invariant, resistant to noise, but not scale invariant.

For the curves in the graphs under changes of blur, JPEG compression and illumination, the rankings of all the descriptors are almost the same. SURF descriptor obtains the lowest recall and highest 1-precision, thus, SURF descriptor is more sensitive to those noises. In addition, SIFT, BRIEF, BRISK, ORB and FREAK descriptors illustrate close recall and 1-precision scores to each other, and it means all of them are robust to the influences of illumination.

The description time complexity of the compared salient point descriptor extraction methods is also statistic in the part. The average time spending on generation of per descriptor based on the dataset provided by Mikolajczyk and Schmid is shown in Table 2. It is clear that binary string descriptors are more efficient than real value descriptors. SIFT descriptor has the highest time consuming, followed by SURF. However, binary string descriptors perform nearly 250 and 30 times faster than SIFT and SURF, respectively. Binary string descriptors are more appropriate for the real-time applications.

Table 2. Comparison of average description time cost

Salient point method	Average extraction time(ms)
SURF+SIFT	7.5
SURF+SURF	0.96
SURF+ORB	0.022
SURF+BRISK	0.023
SURF+FREAK	0.045
SURF+BRIEF	0.023

In applications such as augmented reality, it is very important to not only measure detection accuracy but also to evaluate stability

in the tracking. A concrete example is rendering a cube on a book or a hand. Low stability would be evident in visible vibration in the display of the 3D model and would not result in a positive experience for the viewer. We evaluate stability using a measure of trajectory jitter noise. It is computed as the change in the 3D trajectory which would be used for projecting the augmented reality. In the second experiment, the algorithms were evaluated with regard to jitter noise and accuracy. Moreover, the compared salient point algorithms are combined as follows: SIFT+SIFT, SURF+SURF, ORB+ORB, BRISK+BRISK, SURF+BRIEF and SURF+FREAK. Examples of the used video sequences are shown in Figure 3.

Our trajectory jitter experiments results are shown in Table 3. For a planar object image under affine transformation, it shows that SIFT, SURF, BRISK, and FREAK have better performance with lower variance. However, note there are no jitter results for BRIEF, because it failed to track the object.

Table 3. Comparison of jitter changes under affine changes

Method	Mean jitter	Max jitter	Variance jitter
SIFT+SIFT	0.1566	0.804	0.0113
SURF+SURF	0.3245	1.0541	0.0372
ORB+ORB	4.3226	21.0117	10.8559
BRISK+BRISK	0.5831	6.8654	0.6977
SURF+BRIEF	—	—	—
SURF+FREAK	0.472	1.9839	0.069

The Cambridge Hand Gesture Data set [16] contains different hand shapes and movements. We can see from Table 4 that SIFT had the best performance, while BRISK obtained the lowest score.

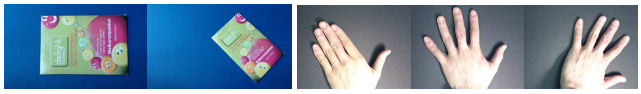

Figure. 3 Examples of planar object and hand data set

Table 4. Comparison on video hand detection accuracy

Salient point method	Average detection accuracy
SIFT+SIFT	68.7%
SURF+SURF	46.3%
ORB+ORB	58.5%
BRISK +BRISK	11.2%
SURF+FREAK	43.3%
SURF+BRIEF	36.9%

4. CONCLUSION

In this paper we presented a comparison of detectors and descriptors on diverse image distortions and also evaluated them on real-time video tracking. No single salient point algorithm was best in all evaluation aspects. The FAST detector had the highest repeatability score than other detectors, moreover and it had the least detection time cost per point. Regarding the criteria of recall-precision, our experiments showed that SIFT, BRISK, and FREAK are the best affine invariant descriptors, and the time complex showed the binary descriptors provide a very efficient description and matching. We also combined detector-descriptor methods for object tracking inside video frames. SIFT and FREAK outperformed the rest in terms of detection accuracy and trajectory jitter. In the future work, we will employ additional metrics and make a further overview on test sets such as the PASCAL and TRECVid benchmarks.

5. REFERENCES

[1] Lowe, D G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision,* 60(2), 91-110.

[2] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3), 346-359.

[3] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. 2010. BRIEF: Binary Robust Independent Elementary Features. *European Conference on Computer vision*, 778–792.

[4] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. 2011. ORB: An Efficient Alternative to SIFT or SURF. *International Conference on Computer Vision*, 2564-2571.

[5] Leutenegger, S., Chli, M., and Siegwart, R. 2011. BRISK: Binary robust invariant scalable keypoints. *International Conference on Computer Vision*, 2548-2555.

[6] Alahi, A., Ortiz, R., and Vandergheynst, P. 2012. FREAK: Fast Retina Keypoint. *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Schmid, C., Mohr, R., and Bauckhage, C. 2000. Evaluation of interest point detectors. *International Journal Computer Vision*, 37(2), 151–172.

[8] Mikolajczyk, K., & Schmid, C. 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1615–1630.

[9] Miksik, O., Mikolajczyk, K. 2012. Evaluation of Local Detectors and Descriptors for Fast Feature Matching. *International Conference on Pattern Recogntion*, 2681-2684.

[10] Gauglitz, S., Höllerer, T., and Turk, M. 2011. Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision*, 335-360.

[11] Rosten, E., and Drummond, T. 2006. Machine learning for high-speed corner detection. *European Conference on Computer Vision*, 430-443.

[12] Agrawal, M., Konolige, K., Blas, M. 2008. Censure: Center Surround Extremas for Realtime Feature Detection and Matching. *European Conference on Computer Vision*, 102–115.

[13] Shi, J., and Tomasi, C. 1994. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition*, 593–600.

[14] Thomee, B., and Lew, M.S. 2012. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 1(2), 71-86.

[15] Jiang, Y., Bhattacharya, S., Chang, S-F., and Shah, M. 2013. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2), 73-101.

[16] Kim, T-K., Wong, S-F., and Cipolla, R. 2007. Tensor Canonical Correlation Analysis for Action Classification, *IEEE Conference on Computer Vision and Pattern Recognition*, 1-8.