FREAK: Fast Retina Keypoint

Alexandre Alahi, Raphael Ortiz, Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

A large number of vision applications rely on matching keypoints across images. The last decade featured an arms-race towards faster and more robust keypoints and association algorithms: Scale Invariant Feature Transform (SIFT)[17], Speed-up Robust Feature (SURF)[4], and more recently Binary Robust Invariant Scalable Keypoints (BRISK)[16] to name a few. These days, the deployment of vision algorithms on smart phones and embedded devices with low memory and computation complexity has even upped the ante: the goal is to make descriptors faster to compute, more compact while remaining robust to scale, rotation and noise.

To best address the current requirements, we propose a novel keypoint descriptor inspired by the human visual system and more precisely the retina, coined Fast Retina Keypoint (FREAK). A cascade of binary strings is computed by efficiently comparing image intensities over a retinal sampling pattern. Our experiments show that FREAKs are in general faster to compute with lower memory load and also more robust than SIFT, SURF or BRISK. They are thus competitive alternatives to existing keypoints in particular for embedded applications.

1. Introduction

Visual correspondence, object matching, and many other vision applications rely on representing images with sparse number of keypoints. A real challenge is to efficiently describe keypoints, *i.e*. image patches, with stable, compact and robust representations invariant to scale, rotation, affine transformation, and noise. The past decades witnessed key players to efficiently describe keypoints and match them.

The most popular descriptor is the histogram of oriented gradient proposed by Lowe [17] to describe the Scale Invariant Feature Transform (SIFT) keypoints. Most of the efforts in the last years was to perform as good as SIFT [14] with lower computational complexity. The Speeded up Robust Feature (SURF) by Bay *et al*. [4] is a good example. It has similar matching rates with much faster performance

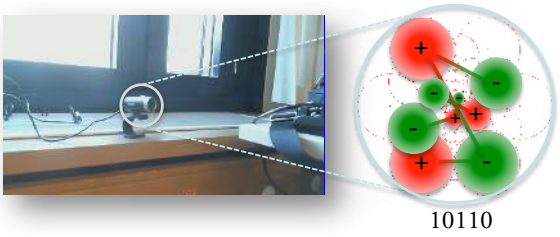


Figure 1: llustration of our FREAK descriptor. A series of Difference of Gaussians (DoG) over a retinal pattern are 1 bit quantized.

by describing keypoints with the responses of few Haar-like filters. In general, Alahi *et al*. show in [2] that a grid of descriptors, similar to SIFT and SURF, is better than a single one to match an image region. Typically, a grid of covariance matrices [30] attains high detection rate but remains computationally too expensive for real-time applications.

The deployment of cameras on every phone coupled with the growing computing power of mobile devices has enabled a new trend: vision algorithms need to run on mobile devices with low computing power and memory capacity. Images obtained by smart phones can be used to perform structure from motion [27], image retrieval [22], or object recognition [15]. As a result, new algorithms are needed where fixed-point operations and low memory load are preferred. The Binary Robust Independent Elementary Feature (BRIEF) [5], the Oriented Fast and Rotated BRIEF (ORB)[26], and the Binary Robust Invariant Scalable Keypoints[16] (BRISK) are good examples. In the next section, we will briefly present these descriptors. Their stimulating contribution is that a binary string obtained by simply comparing pairs of image intensities can efficiently describe a keypoint, *i.e*. an image patch. However, several problems remain: how to efficiently select the ideal pairs within an image patch? How to match them? Interestingly, such trend is inline with the models of the nature to describe complex observations with simple rules. We propose to address such unknowns by designing a descriptor inspired by the Human Visual System, and more precisely the retina. We propose the Fast Retina Keypoint (FREAK) as a fast,

compact and robust keypoint descriptor. A cascade of binary strings is computed by efficiently comparing pairs of image intensities over a retinal sampling pattern. Interestingly, selecting pairs to reduce the dimensionality of the descriptor yields a highly structured pattern that mimics the saccadic search of the human eyes.

2. Related work

Keypoint descriptors are often coupled with their detection. Tuytelaar *et al*. in [29] and Gauglitz *et al*. in [11] presented a detailed survey. We briefly present state-of-the-art detectors and mainly focus on descriptors.

2.1. Keypoint detectors

A first solution is to consider corners as keypoints. Harris and Stephen in [12] proposed the Harris corner detector. Mikolajczyk and Schmid made it scale invariant in [20]. Another solution is to use local extrema of the responses of certain filters as potential keypoints. Lowe in [17] filtered the image with differences of Gaussians. Bay *et al*. in [4] used a Fast Hessian detector. Agrawal *et al*. in [1] proposed simplified center-surround filters to approximate the Laplacian.. Ebrahimi and Mayol-Cuevas in [7] accelerated the process by skipping the computation of the filter response if the response for the previous pixel is very low. Rostenand and Drummond proposed in [25] the FAST criterion for corner detection, improved by Mair *et al*. in [18] with their AGAST detector. The latter is a fast algorithm to locate keypoints. The detector used in BRISK by Leutenegger *et al*. in [16] is a multi-scale AGAST. They search for maxima in scale-space using the FAST score as a measure of saliency. We use the same detector for our evaluation of FREAK.

2.2. SIFT-like descriptors

Once keypoints are located, we are interested in describing the image patch with a robust feature vector. The most well-known descriptor is SIFT [17]. A 128-dimensional vector is obtained from a grid of histograms of oriented gradient. Its high descriptive power and robustness to illumination change have ranked it as the reference keypoint descriptor for the past decade. A family of SIFT-like descriptor has emerged in the past years. The PCA-SIFT [14] reduces the description vector from 128 to 36 dimension using principal component analysis. The matching time is reduced, but the time to build the descriptor is increased leading to a small gain in speed and a loss of distinctiveness. The GLOH descriptor [21] is an extension of the SIFT descriptor that is more distinctive, but also more expensive to compute. The robustness to change of viewpoint is improved in [31] by simulating multiple deformations to the descriptive patch. Good compromises between performances and the number of simulated patches lead to an algorithm twice slower than SIFT. Ambai and Yoshida proposed a Compact And Real-time Descriptors (CARD) in [3] to extract the histogram of oriented gradient from the grid binning of SIFT or the log-polar binning of GLOH. The computation of the histograms is simplified by using lookup tables.

One of the widely used keypoints at the moment is clearly SURF [4]. It has similar matching performances as SIFT but is much faster. It also relies on local gradient histograms. The Haar-wavelet responses are efficiently computed with integral images leading to 64 or 128-dimensional vectors. However, the dimensionality of the feature vector is still too high for large-scale applications such as image retrieval or 3D reconstruction. Often, Principal Component Analysis (PCA), or hashing functions are used to reduce the dimensionality of the descriptors [24]. Such steps involve time-consuming computation and hence affect the real-time performance.

2.3. Binary descriptors

Calonder *et al*. in [5] showed that it is possible to shortcut the dimensionality reduction step by directly building a short binary descriptor in which each bits are independent, called BRIEF. A clear advantage of binary descriptors is that the Hamming distance (bitwise XOR followed by a bit count) can replace the usual Euclidean distance. The descriptor vector is obtained by comparing the intensity of 512 pairs of pixels after applying a Gaussian smoothing to reduce the noise sensitivity. The positions of the pixels are pre-selected randomly according to a Gaussian distribution around the patch center. The obtained descriptor is not invariant to scale and rotation changes unless coupled with detector providing it. Calonder *et al*. also highlighted in their work that usually orientation detection reduces the recognition rate and should therefore be avoided when it is not required by the target application. Rublee *et al*. in [26] proposed the Oriented Fast and Rotated BRIEF (ORB) descriptor. Their binary descriptor is invariant to rotation and robust to noise. Similarly, Leutenegger *et al*. in [16] proposed a binary descriptor invariant to scale and rotation called BRISK. To build the descriptor bit-stream, a limited number of points in a specific sampling pattern is used. Each point contributes to many pairs. The pairs are divided in short-distance and long-distance subsets. The long-distance subset is used to estimate the direction of the keypoint while the short-distance subset is used to build binary descriptor after rotating the sampling pattern.

In Section 5, we compare our proposed FREAK descriptor with the above presented descriptors. But first, we present a possible intuition on why these trendy binary descriptors can work based on the study of the human retina.

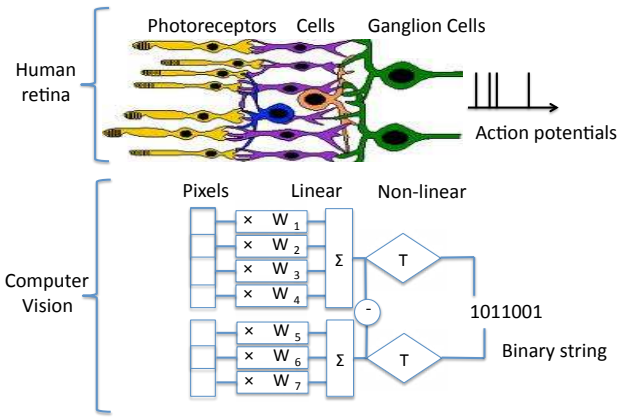
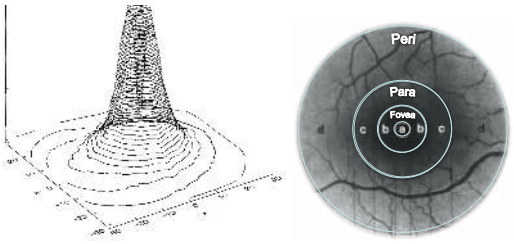


Figure 2: From human retina to computer vision: the biological pathways leading to action potentials is emulated by simple binary tests over pixel regions. [Upper part of the image is a courtesy of the book Avian Visual Cognition by R: Cook].



(a) Density of ganglion cells over the retina [10]. (b) Retina areas [13]

Figure 3: Illustration of the distribution of ganglion cells over the retina. The density is clustered into four areas: (a) the foveola, (b) fovea, (c) parafoveal, and (d) perifoveal.

3. Human retina

3.1. Motivations

In the presented literature, we have seen that recent progress in image representation has shown that simple intensity comparison of several pairs of pixels can be good enough to describe and match image patches [5, 26, 16]. However, there exist some open interrogations on the ideal selection of pairs. How should we sample them and compare them? How to be robust to noise? Should we smooth with a single Gaussian kernel? In this work, we show how to gain performance by selecting a solution inspired by the human retina, while enforcing low computational complexity.

Neuroscience has made lots of progress in understanding the visual system and how images are transmitted to the brain [8]. It is believed that the human retina extracts details from images using Difference of Gaussians (DoG) of various sizes and encodes such differences with action potentials. The topology of the retina plays an important role. We propose to mimic the same strategy to design our image descriptor.

3.2. Analogy: from retinal photoreceptors to pixels

The topology and spatial encoding of the retina is quite fascinating. First, several photoreceptors influence a ganglion cell. The region where light influences the response of a ganglion cell is the receptive field. Its size and dendritic field increases with radial distance from the foveola (Figure 3). The spatial distribution of ganglion cells reduces exponentially with the distance to the foveal. They are segmented into four areas: foveal, fovea, parafoveal, and perifoveal. Each area plays an interesting role in the process of detecting and recognizing objects since higher resolution is captured in the fovea whereas a low acuity image is formed in the perifoveal. One can interpret the decrease of resolution as a body resource optimization. Let us now turn these insights into an actual keypoint descriptor. Figure 2 presents the proposed analogy.

4. FREAK

4.1. Retinal sampling pattern

Many sampling grids are possible to compare pairs of pixel intensities. BRIEF and ORB use random pairs. BRISK uses a circular pattern where points are equally spaced on circles concentric, similar to DAISY [28]. We propose to use the retinal sampling grid which is also circular with the difference of having higher density of points near the center. The density of points drops exponentially as can be seen in Figure 3. .

Each sample point needs to be smoothed to be less sensitive to noise. BRIEF and ORB use the same kernel for all points in the patch. To match the retina model, we use different kernels size for every sample points similar to BRISK. The difference with BRISK is the exponential change in size and the overlapping receptive fields. Figure 4 illustrates the topology of the receptive fields. Each circle represents the standard deviations of the Gaussian kernels applied to the corresponding sampling points.

We have experimentally observed that changing the size of the Gaussian kernels with respect to the log-polar retinal pattern leads to better performance. In addition, overlapping the receptive fields also increases the performance. A possible reason is that with the presented overlap in Figure 4, more information is captured. We add redundancy that brings more discriminative power. Let's consider the intensities I_i measured at the receptive fields A, B, and C where:

$$I_A > I_B, I_B > I_C, \text{and } I_A > I_C. \quad (1)$$

If the fields do not have overlap, then the last test $I_A > I_C$ is not adding any discriminant information. However,

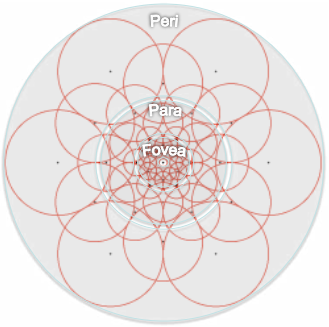


Figure 4: Illustration of the FREAK sampling pattern similar to the retinal ganglion cells distribution with their corresponding receptive fields. Each circle represents a receptive field where the image is smoothed with its corresponding Gaussian kernel.

if the fields overlap, partially new information can be encoded. In general, adding redundancy allow us to use less receptive fields which is a known strategy employed in compressed sensing or dictionary learning [6]. According to Olshausen and Field in [23], such redundancy also exists in the receptive fields of the retina.

4.2. Coarse-to-fine descriptor

We construct our binary descriptor F by thresholding the difference between pairs of receptive fields with their corresponding Gaussian kernel. In other words, F is a binary string formed by a sequence of one-bit Difference of Gaussians (DoG):

$$F = \sum_{0 \leq a < N} 2^a T(P_a), \tag{2}$$

where P_a is a pair of receptive fields, N is the desired size of the descriptor, and

$$T(P_a) = \begin{cases} 1 & \text{if } (I(P_a^{r_1}) - I(P_a^{r_2}) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

with $I(P_a^{r_1})$ is the smoothed intensity of the first receptive field of the pair P_a.

With few dozen of receptive fields, thousands of pairs are possible leading to a large descriptor. However, many of the pairs might not be useful to efficiently describe an image. A possible strategy can be to select pairs given their spatial distance similar to BRISK. However, the selected pairs can be highly correlated and not discriminant. Consequently, we run an algorithm similar to ORB [26] to learn the best pairs from training data:

1. We create a matrix D of nearly fifty thousands of extracted keypoints. Each row corresponds to a keypoint represented with its large descriptor made of all possible pairs in the retina sampling pattern illustrated in Figure 4. We use 43 receptive fields leading to approximately one thousand pairs.

2. We compute the mean of each column. In order to have a discriminant feature, high variance is desired. A mean of 0.5 leads to the highest variance of a binary distribution.

3. We order the columns with respect to the highest variance.

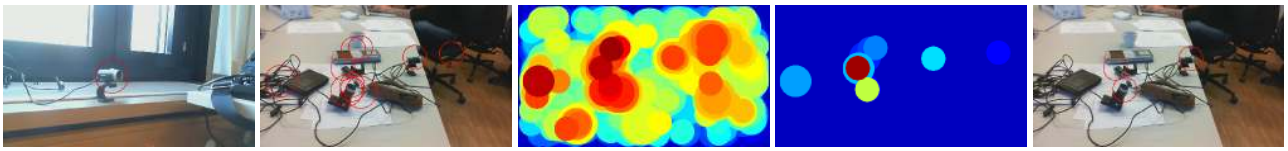
4. We keep the best column (mean of 0.5) and iteratively add remaining columns having low correlation with the selected columns.

Strikingly, there is a structure in the selected pairs. A coarse-to-fine ordering of the difference of Gaussians (the pairs) is automatically prefered. Figure 5 illustrates the pairs selected by grouping them into four clusters (128 pairs per group). We experimentally observed that the first 512 pairs are the most relevant and adding more pairs is not increasing the performance. A symmetric scheme is captured due to the orientation of the pattern along the global gradient. Interestingly, the first cluster involves mainly peripheral receptive fields whereas the last ones implicates highly centered fields. It appears to be reminiscent of the behavior of the human eye. We first use the perifoveal receptive fields to estimate the location of an object of interest. Then, the validation is performed with the more densely distributed receptive fields in the fovea area. Although the used feature selection algorithm is a heuristic, it seems to match our understanding of the model of the human retina. Our matching step takes advantage of the coarse-to-fine structured of FREAK descriptor. Note that in the last decades, coarse-to-fine strategy has often been explored in detecting and matching objects [9, 2].

4.3. Saccadic search

Humans do not look at a scene in fixed steadiness. Their eyes move around with discontinuous individual movements called saccades. The presented cells topology in the retina is one reason for such movements. As explained previously, the fovea captures high-resolution information thanks to its high-density photoreceptors. Hence, it plays a critical role in recognizing and matching objects. The perifoveal area captures less detailed information, *i.e.* low-frequency observation. Consequently, they are used to compile first estimates of the locations of our objects of interest.

We propose to mimic the saccadic search by parsing our descriptor in several steps. We start by searching with the first 16 bytes of the FREAK descriptor representing coarse information. If the distance is smaller than a threshold, we further continue the comparison with the next bytes to analyze finer information. As a result, a cascade of comparisons is performed accelerating even further the matching



(a) An object of interest described by 1 FREAK (b) Best matches after 1^{st} cascade (c) Distance map with first cascade (in color jet) (d) Distance map with last cascade (in color jet) (e) Matched objects after last cascade

Figure 6: Illustration of the cascade approach.

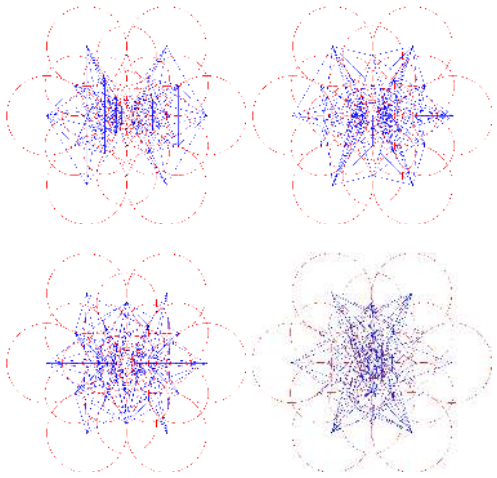


Figure 5: Illustration of the coarse-to-fine analysis. The first cluster involves mainly perifoveal receptive fields and the last ones fovea.

step. More than 90% of the candidates are discarded with the first 16 bytes of our FREAK descriptor. Note that we have chosen 16 bytes for the first cascade to match hardware requirements. To compare 1 byte or 16 bytes is almost equivalent with Single Instruction and Multiple Data (SIMD) instructions on Intel processors since operations are performed in parallel.

We illustrate the saccadic search in Figure 6. For visualization purposes, we describe an object of interest with a single FREAK descriptor of the size of its bounding circle (Figure 6 (a)). Then, we search for the same object in a new image. All candidate image regions are also described with a single descriptor of the size of the candidate region. The first cascade (first 16 bytes) discards many candidates and select very few of them to compare with the remaining bytes. In Figure 6 (e), the last cascade has correctly selected the locations of our object of interest despite the changes of illuminations and viewpoints.

4.4. Orientation

In order to estimate the rotation of our keypoint, we sum the estimated local gradients over selected pairs similar to BRISK [16]. The latter is using long pairs to compute the global orientation whereas we mainly select pairs with symmetric receptive fields with respect to the center (see Figure 7.

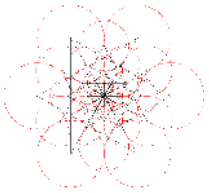


Figure 7: Illustration of the pairs selected to computed the orientation.

Let G be the set of all the pairs used to compute the local gradients:

$$O = \frac{1}{M} \sum_{P_o \in G} (I(P_o^{r_1}) - I(P_o^{r_2})) \frac{P_o^{r_1} - P_o^{r_2}}{\|P_o^{r_1} - P_o^{r_2}\|}, \quad (3)$$

where M is the number of pairs in G and $P_o^{r_i}$ is the 2D vector of the spatial coordinates of the center of receptive field.

We select 45 pairs as opposed to the few hundreds pairs of BRISK. In addition, our retinal pattern has larger receptive fields in the perifoveal area than BRISK allowing more error in the orientation estimation. We therefore discretize the space of orientations in much bigger steps leading to more than 5 times smaller memory load (about 7 MB against 40 MB).

5. Performance evaluation

The evaluation of keypoints has been exhaustively studied in the literature. Two testing environments are used: first, the well-known dataset introduced by Mikolajczyk and Schmid [21]. We present the recall (number of correct matches / number of correspondences) vs 1-precision curve

Time per keypoint	SIFT	SURF	BRISK	FREAK
Description in [ms]	2.5	1.4	0.031	0.018
Matching time in [ns]	1014	566	36	25

Table 1: Computation time on 800x600 images where approximately 1500 keypoints are detected per image. The computation times correspond to the description and matching of all keypoints.

(number of false matches / number of matches). We also evaluate FREAK on a more task specific framework similar to the one proposed on-line[1]. The "graf1" image from the previous dataset is transformed to precisely evaluate the rotation, change of scale, change of viewpoint, blur (gaussian), and brightness change. We can continuously measure the impact of an image deformation, which is less visible in the first testing environment. Figures 8 to 10 present the quantitative results.

The performances of the descriptors are highly related to the combination detector/descriptor. Some descriptors are more discriminant for blobs than corners. Nevertheless, we noticed that the global ranking of their matching performance remain the same regardless of the selected detector. As a result, we present our tests using the multi-scale AGAST detector introduced by BRISK.

Both testing environments rank FREAK as the most robust to all the tested image deformation. Surprisingly, SIFT is the worst descriptor in the first testing environment (8) similar to what has been shown in BRISK [16]. SIFT extracts several descriptors when the estimated orientation has multiple good candidates. Therefore, the number of possible correspondences is over estimated. Nevertheless, the second testing environment (9) is not affected by the additional descriptors created with SIFT. Consequently, SIFT is more competitive and often ranks second. Figure 10 present the performance of descriptors that are not invariant to rotation and scale invariant. We compare FREAK while disabling its scale and orientation invariance. It is either as good as other descriptor or slightly better.

Table 1 compares the computation time of the scale and rotation invariant descriptors. All algorithms are running on an Intel Duo core of 2.2 GHZ using a single core. FREAK is even faster than BRISK although the latter is two orders of magnitude faster than SIFT and SURF.

6. Conclusions

We have presented a retina-inspired keypoint descriptor to enhance the performance of current image descriptors. It outperforms recent state-of-the-art keypoint descriptors while remaining simple (faster with lower memory load). We do not claim any biological significance but find it remarkable that the used learning stage to identify the most relevant Difference of Gaussians could match one possible understanding of the resource optimization of the human visual system. In fact, as a future work, we want to investigate more on the selection of such relevant pairs for high level applications such as object recognition.

[1] http://computer-vision-talks.com/2011/08/feature-descriptor-comparison-report/

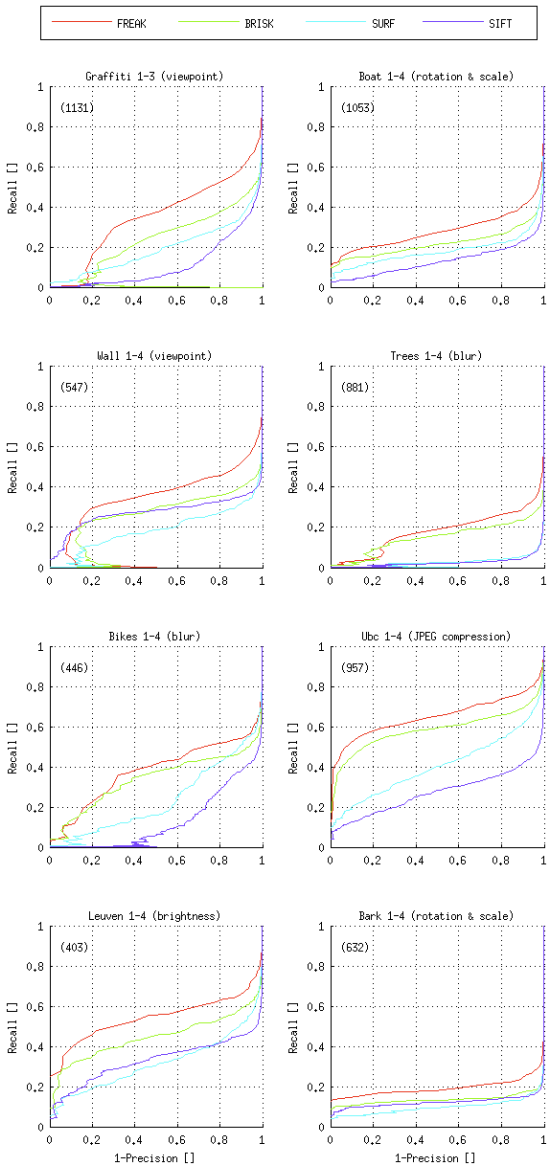


Figure 8: Performance evaluation on the dataset introduced by Mikolajczyk and Schmid [21].

References

[1] M. Agrawal, K. Konolige, and M. Blas. Censure: Center surround extremas for realtime feature detection and matching. *Computer Vision–ECCV 2008*, pages 102–115, 2008. 2

[2] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt. Cascade of descriptors to detect and track objects across any net-

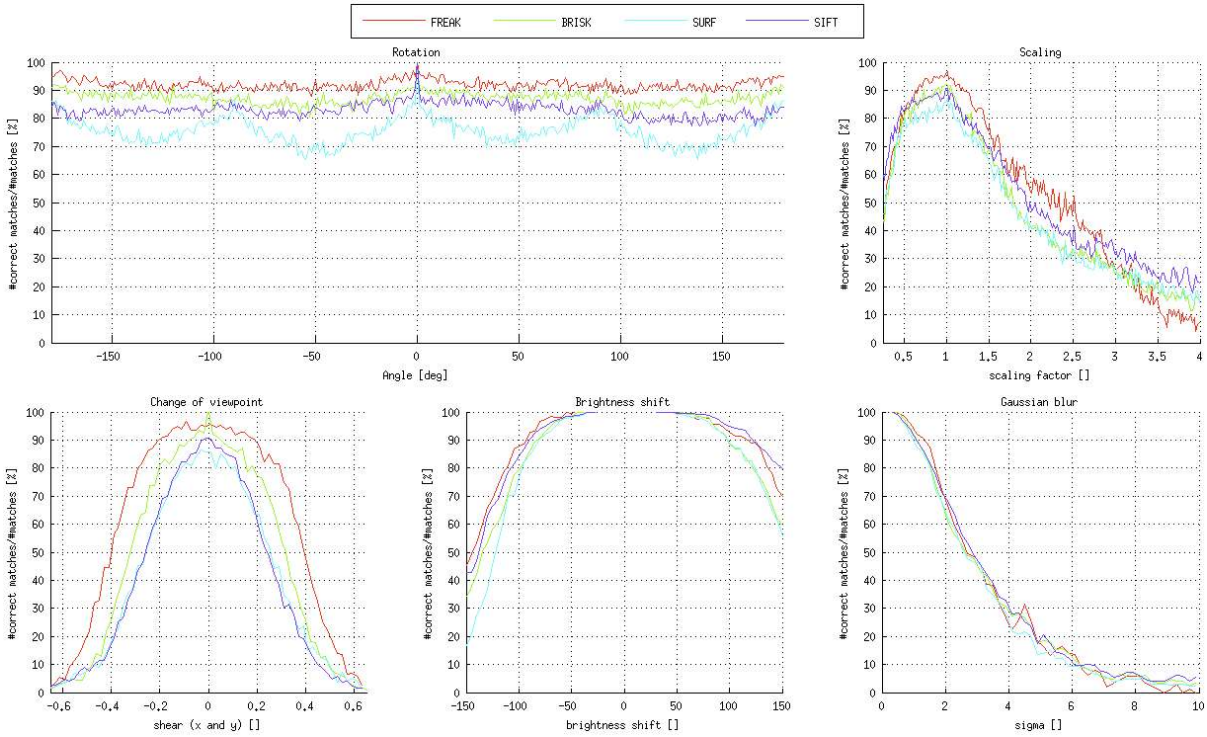


Figure 9: Global evaluation on the second testing environment. The view point changes are simulated by an affine matrix transformation where the shear parameter goes from -0.65 to 0.65 in both x and y directions.

work of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010. 1, 4

[3] M. Ambai and Y. Yoshida. Card: Compact and real-time descriptors. In *Computer Vision, 2011 IEEE Computer Society Conference on*. IEEE, 2011. 2

[4] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006. 1, 2

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, pages 778–792, 2010. 1, 2, 3

[6] E. Candes, Y. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011. 4

[7] M. Ebrahimi and W. Mayol-Cuevas. Susure: Speeded up surround extrema feature detector and descriptor for realtime applications. 2009. 2

[8] G. Field, J. Gauthier, A. Sher, M. Greschner, T. Machado, L. Jepson, J. Shlens, D. Gunning, K. Mathieson, W. Dabrowski, et al. Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, 467(7316):673–677, 2010. 3

[9] F. Fleuret and D. Geman. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41(1):85–107, 2001. 4

[10] D. Garway-Heath, J. Caprioli, F. Fitzke, and R. Hitchings. Scaling the hill of vision: the physiological relationship between light sensitivity and ganglion cell numbers. *Investigative Ophthalmology & Visual Science*, 41(7):1774–1782, 2000. 3

[11] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision*, pages 1–26, 2011. 2

[12] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988. 2

[13] M. Hogan and J. JA Weddell. Histology of the human eye: an atlas and textbook. 1971. 3

[14] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. 2004. 1, 2

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. Ieee, 2006. 1

[16] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. 2011. 1, 2, 3, 5, 6

[17] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee, 1999. 1, 2

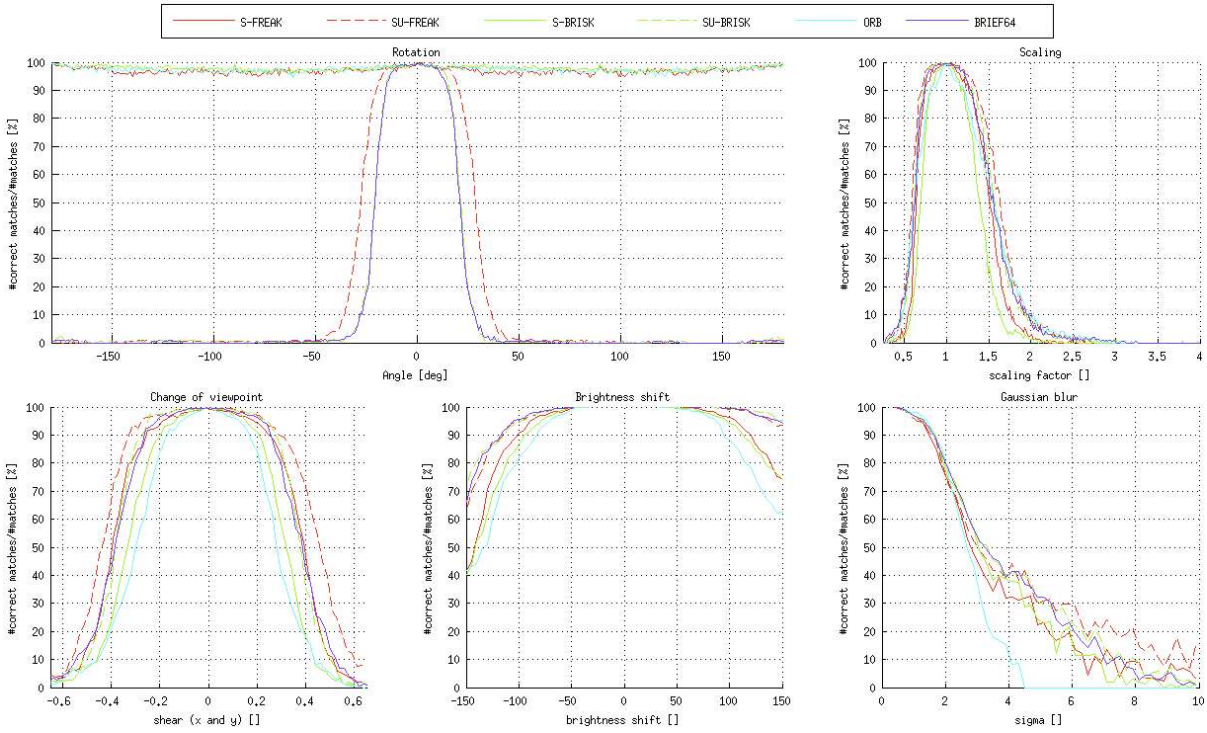


Figure 10: Global evaluation on the second testing environment: (S: single scale; U: unrotated) BRIEF64, S-BRISK, SU-BRISK, S-FREAK, SU-FREAK descriptors are extracted on keypoints detected by FAST.

[18] E. Mair, G. Hager, D. Burschka, M. Suppa, and G. Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. *Computer Vision–ECCV 2010*, pages 183–196, 2010. 2

[19] E. Marc and H. Stuart. Lateral information processing by spiking neurons: A theoretical model of the neural correlate of consciousness. *Computational Intelligence and Neuroscience*, 2011, 2011.

[20] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. 2001. 2

[21] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, pages 1615–1630, 2005. 2, 5, 6

[22] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE, 2006. 1

[23] B. Olshausen and D. Field. What is the other 85% of v1 doing. *Problems in Systems Neuroscience*, pages 182–211, 2004. 4

[24] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[25] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. *Computer Vision–ECCV 2006*, pages 430–443, 2006. 2

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. 2011. 1, 2, 3, 4

[27] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006. 1

[28] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010. 3

[29] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008. 2

[30] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *Computer Vision–ECCV 2006*, pages 589–600, 2006. 1

[31] G. Yu and J. Morel. A fully affine invariant image comparison method. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1597–1600. IEEE, 2009. 2