RIFF: Retina-inspired Invariant Fast Feature Descriptor

Song Wu Michael S. Lew
LIACS Media Lab, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
{s.wu, m.s.lew}@liacs.leidenuniv.nl

ABSTRACT
We present the Retina-inspired Invariant Fast Feature, RIFF, which was designed for invariance to scale, rotation, and affine image deformations. The feature descriptor is based on pair-wise comparisons over a sampling pattern loosely based on the human retina and introduces a method for improving accuracy by maximizing the discriminatory power of the point set. A performance evaluation with regard to bag of words based image retrieval on several well-known international datasets demonstrates that the RIFF descriptor has competitive performance to the state-of-the-art descriptors (*e.g.* SIFT, SURF, BRISK, and FREAK).

Categories and Subject Descriptors
Computing methodologies, computer vision

General Terms
Algorithms, Performance, Experimentation

Keywords
Salient point descriptor; bag of words; image copy detection

1. INTRODUCTION
Efficiently establishing the correspondences between images is very useful for numerous applications of computer vision and content-based retrieval, such as content-based image retrieval, image classification, object tracking, and panorama stitching. Salient point methods are a leading approach which has been proven to be effective in many real world applications.

In using salient points, one typically needs a detector and a descriptor. Detectors find the locations (i.e. blob, region, and point) in images which typically are in some way informative. The descriptor gives a model or representation of the local image region. Prior research of salient points has focused on high repeatability detector and robustness to scale and rotation [1].

The SIFT descriptor [2] is the most popular salient point approach. It computes the Difference-of-Gaussian (DoG) operator in the Gaussian scale space, and assigns orientation and descriptor to each salient point based on local gradient histogram. The SURF [3] salient points detector makes use of a box-filter to achieve efficient extrema detection in scale space and it performs well with respect to the criteria of repeatability. The SURF descriptor of each detected salient point is calculated through summing Haar-wavelet responses in the defined region after orientation alignment. Recent binary string descriptors such as BRIEF, ORB, BRISK, and FREAK were proposed with specific advantages such as low memory requirements as well as efficiently matching via hamming distance (bitwise XOR followed by a bit count). BRIEF [4] first uses Gaussian smoothing on the selected image patch, and creates a binary string descriptor via the intensity comparison of randomly sampled pixel-pairs around the patch center. ORB [5] employs the most efficient FAST [6] detector to determine the salient points in different layers of an image pyramid, creates an orientation for each point by the intensity centroid algorithm. It forms the binary string descriptor based on BRIEF and effectively improves the sensitivity to image rotation and scale. BRISK [7] applies FAST score as a measure to determine the extreme points in the image scale pyramid, and generates the descriptor by comparing pair-wise intensities over a decreasing density circular sampling pattern. FREAK [8] also selects pairs of pixels over a decreasing density circular sampling pattern loosely inspired by the retina and then compares their intensities to form a binary vector. Both BRISK and FREAK use the sum of local gradients of selected pairs to estimate the orientation.

The recent wave of salient point detectors each have specific strengths. Some are best for scale changes; others for speed; others for memory requirements. Our goal was to design a detector which was optimized for affine transformations including rotation and scaling. In this paper, we propose a novel discriminate salient point real value descriptor named "RIFF" because the sampling pattern is inspired by the distribution of cones (color vision) in the human eye. The main contributions of this paper are as follows: first, we describe a salient point descriptor which outperforms current methods regarding affine transformations. Moreover, we proposed a measure to rank the generated salient point descriptors so that unstable points will be rejected and the discriminatory power of the set of descriptors will be improved. This is useful for speeding up the process of indexing and matching among large scale descriptors and increasing accuracy.

The rest of the paper is organized as follows: In Section 2, we present the generation of our RIFF local feature descriptor. In Section 3, we describe the datasets and evaluation criterion in the experiment. The performance result of the proposed descriptor compared with current state-of-the-art descriptors are shown in Section 4, and conclusions are given in Section 5.

2. DISCRIMINATE RIFF SALIENT POINT DESCRIPTOR METHOD

2.1 Retina Sampling Pattern Review
The retina sampling pattern is based on topology of human retina from the neuroscience research, which reveals that the spatial

MM '14, November 3–7 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2654994

distribution density of cone cells in the retina decreases exponentially with the distance measure to the center of fovea. Moreover, it is believed that the signal of image passes through from cone cells to ganglion cells and the receptive field of each ganglion cell using Difference of Gaussians (DoG) model with various sizes and encodes such differences into action potentials. Our approach employed the similar retina sampling pattern, which place different size of blocks at the defined location in the pattern. The illustration of the cones density can be seen in Figure 1.

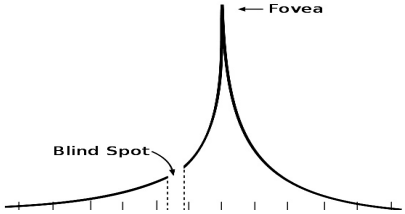
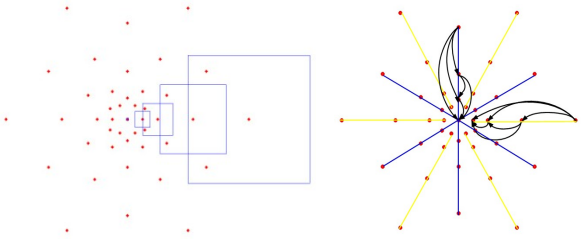


Figure 1. Illustration of the density distribution of cones in the human retina

Inspired by recent work with decreasing circular polar densities in diverse applications from stereo matching to object recognition [7, 8, 9], the sampling pattern for RIFF in 2D decreases exponentially as shown in Figure 2 (a).



(a) (b)

Figure 2. (a) The 2D decreasing exponential polar sampling pattern for RIFF with N=43 points: the red point denotes the sampling point location, the blue rectangle represents a receptive field, and the size of rectangle corresponds to its Gaussian kernel which used to smooth the intensity values at the sampling points. (b) The pre-defined pair-wise point comparisons on RIFF for 2 of the 12 axes.

2.2 Descriptor Generation

2.2.1 Orientation estimation

Given a set of salient points in an image (detected by the salient point detector), we first position and scale the retina sampling pattern according to the location and scale information (this is computed by the detector) for each specified point, and then calculate an orientation for them.

The popular approach for estimating the orientation angle comes from basic geometry which is to estimate the orientation using local gradients: Δy and Δx and then determine the angle from the arctangent of ($\Delta y/\Delta x$) (for details see FREAK [8]). We also estimate the local gradients by pair-wise differences between equidistant points from the center.

2.2.2 Descriptor building

The procedure of RIFF descriptor generation is different from previous salient point approaches such as BRIEF, ORB, BRISK and FREAK which compare the pixel-pairs intensity in the sampling pattern to generate a binary string feature. Our approach first constructs a structure in the retina sampling pattern rotated by the estimated orientation θ. Let V= $[v_1,...v_i,...v_d]$ represent a feature vector of a salient point, v_i is a float value obtained by calculating the difference of Gaussian smoothed image intensities of pre-defined pairs over the structure. We defined 6 pair-wise comparisons on each of the 12 axes from the center which resulted in the dimension of the descriptor d as 72. For clarity, we displayed in Figure 2(b) 1 of 6 pair-wise comparisons on the blue axes and 1 of 6 on the yellow axes where each black curve denotes one pair comparison. Since we place a block at each sampling point, the integral image (summed area tables) was used for computational efficiency. Compared with the binary string features, it was not necessary for RIFF to compare the intensity of all possible N*(N-1)/2 sampling pairs, moreover, the dimension of RIFF is smaller than SIFT, which can improve the speed of indexing and matching.

2.2.3 Discriminatory Strategy

Even though location, scale and orientation have been estimated, current salient point detectors have difficulty with affine viewpoint changes such as in Figure 3. We conducted a small internal study which revealed that local ambiguities (nearby salient points with similar feature descriptors) are often the cause of those matching errors.

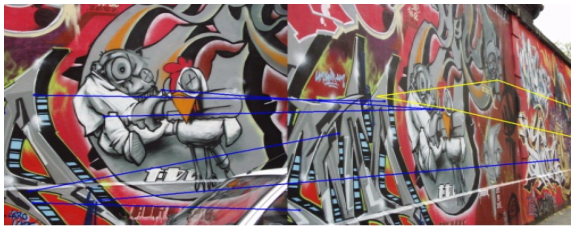


Figure 3. Matches (blue lines) from using SIFT (OpenCV) salient point approach.

Thus, our goal was to reduce local ambiguity or increase local distinctiveness by eliminating salient points which have similar salient points nearby. We implemented this process by using a ranking scheme to identify stable local features as described next. Consider a set of salient point descriptors $\{f_i,\ i=1,2,...,M\}$, a salient point p in the image I and its feature is f_p. The discriminatory score of the feature is defined according to the measure of similarity when compared to its K nearest neighbors in the image.

$$D_p(p \in I) = \sum_{j=1}^{K} \left\| f_p - f_j \right\|_2 \quad (1)$$

$\|\cdot\|$ denotes the Euclidean distance. Intuitively, the higher discriminatory score demonstrates that the feature of point p is more distinctive than other points. The parameter K is set to 2 in the experiment. Furthermore, we use an exponential function in order to emphasize the discriminate score:

$$D'_p(p \in I) = \exp(-\lambda \cdot \left| D_p \right|) \quad (2)$$

$|\cdot|$ denotes the normalization of D_p (in the range [0, 1]), λ is a weight of discriminate score and it is set to 6 in the experiment. We note that after the above process, the smaller D'_p score correlates to more distinctive feature points, so we can sort these scores and define a threshold to filter those unstable salient points. The final set is a smaller number of discriminative features which have are more robust to various image transformations, while

(a) RIFF (b) SIFT (c) SURF (d) FREAK

Figure 5. Illustration of descriptors matching, RIFF as compared to SIFT, SURF, and FREAK on challenge affine object detection (graffiti 1-5 proposed by Mikolajczyk and Schmid[1]).

reducing required subsequent processing, *e.g.,* descriptor indexing as well as dictionary learning in large scale image applications.

3. DATASETS AND EVALUATION CRITERIA

We evaluated our proposed RIFF descriptor with state-of-art local descriptors on three well known benchmark datasets in terms of bag of words based image retrieval. The Nearest Neighbor Distance Ratio (NNDR) is used as the matching strategy to define the similar descriptors between two images. PASCAL VOC 2012 dataset [10], Caltech 256 [11], as well as MIRFLICKER 1M consisting of one million images [12] were used to evaluate the performance on large scale similar (distorted/transformed duplicate) image detection. Additional experiments (stability, recall and precision, etc.) which did not fit space limitations can be found here: http://press.liacs.nl/researchdownloads/. The two descriptors are viewed as a correspondence if $\|D_A\text{-}D_B\|/\|D_A\text{-}D_C\|<threshold$, where D_B is the first and D_C is the second nearest neighbor of D_A.

We used mAP (mean Average Precision) as a criterion for the evaluation of detection accuracy. The transformed duplicates categories generated for the test mainly include: cropping, content noise, image blur, image compression: JPEG compression 5%-95%, rotation: 30-360 deg., scale: 20-200% and affine transformation: rotation + scale + 60 deg. 3D perspective distortion (Figure 4). 1000 images in each dataset were chosen as queries, and each query image corresponds to 80 duplicates in the evaluation.

Figure 4. Types of image transforms: original (left), rotation and scale (middle), affine - 3D pan rotation (right).

4. EXPERIMENTS AND RESULTS

In the performance evaluation of our proposed RIFF descriptor, we used the SURF salient point detector for locating salient points due to its high repeatability [7]. RIFF was programmed in C++ and the descriptor and datasets can be downloaded from http://press.liacs.nl/researchdownloads/.

4.1 Descriptors Matching

In this part, we set the value of *threshold* in the NNDR 0.75, and the homography between two compared images is estimated by the RANSAC algorithm. In preliminary tests, RIFF exhibited competitive performance in the cases of affine image transformations in comparison to the popular SIFT, SURF, and recent FREAK descriptors as shown in Figure 5.

Time cost and memory space requirement were also evaluated in this part. As summarized in Table 1, binary string features (BRIEF, ORB, BRISK and FREAK) were more computationally efficient and space effective (using an Intel Core i7 Processor (2.67GHz), 12GB of RAM)

Table 1. Comparison of descriptors in terms of extraction time and memory space requirement (5000 descriptors)

Methods	Time cost	memory requirement
SURF+SIFT	5.5 Seconds	2.44M
SURF+SURF	0.3 Seconds	1.22M
SURF+BRIEF	0.044 Seconds	0.15M
SURF+ORB	0.045 Seconds	0.15M
SURF+BRISK	0.058 Seconds	0.3M
SURF+FREAK	0.1 Seconds	0.3M
SURF+RIFF	0.38 Seconds	1.37M

4.2 Image Retrieval Experiments

In this section, we evaluated the proposed RIFF feature descriptor in the area of large scale similar image detection. The bag of visual words model was used. The visual vocabulary was first trained based on the extracted descriptors from PASCAL VOC

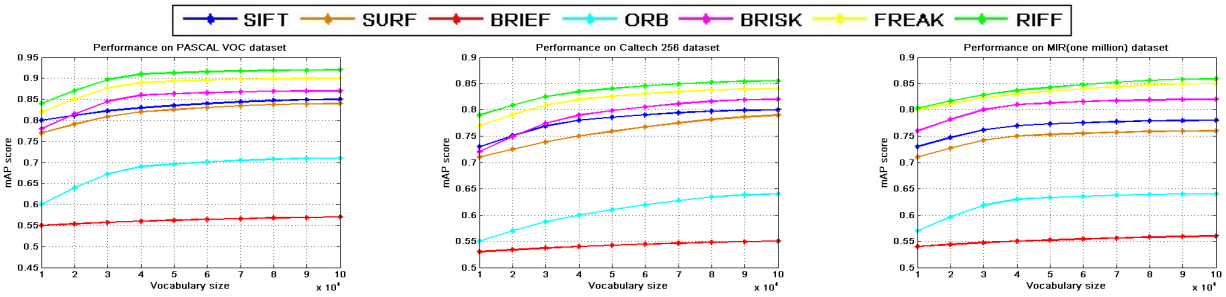


Figure 6. Detection accuracy on three datasets (PASCAL VOC, Caltech 256 and MIRFLICKER).

dataset (with dimensions from 10K to 100K). We used the ANN search to speed up the vocabulary generation. Due to the different properties of real value descriptors and binary string descriptors, the ANN search was based on KD-tree index and multi-probe LSH index respectively [13]. Once the visual vocabulary was generated, descriptors of the images were encoded into a histogram according to the occurrence frequency of each visual word together with the *tf-idf* weighting scheme [14]. The cosine distance measure was adopted to estimate similarity of two images represented by visual words. The compared visual vocabularies were generated by different types of descriptors and the detection accuracy was measured by the mAP score on three datasets. In the following step, we analyzed the final rank and distribution of each category's transformed duplicates on the MIRFLICKER one million dataset.

Overall, RIFF outperformed the other descriptors on the PASCAL VOC, Caltech 256 and MIRFLICKR-1M datasets as shown in Figure 6. Regarding specific transformations as shown in Figure 7, RIFF had the best performance on the distortions related to scale, rotation, and affine transformations. It had average performance on blurring and had competitive performance on the rest of the transformations.

RIFF was roughly 14 times faster than SIFT which makes it amenable to real-time applications. It is slower (and requires more memory) than FREAK, however, it has shown significantly higher accuracy regarding affine transformations as displayed in Figure 5 and Figure 7.

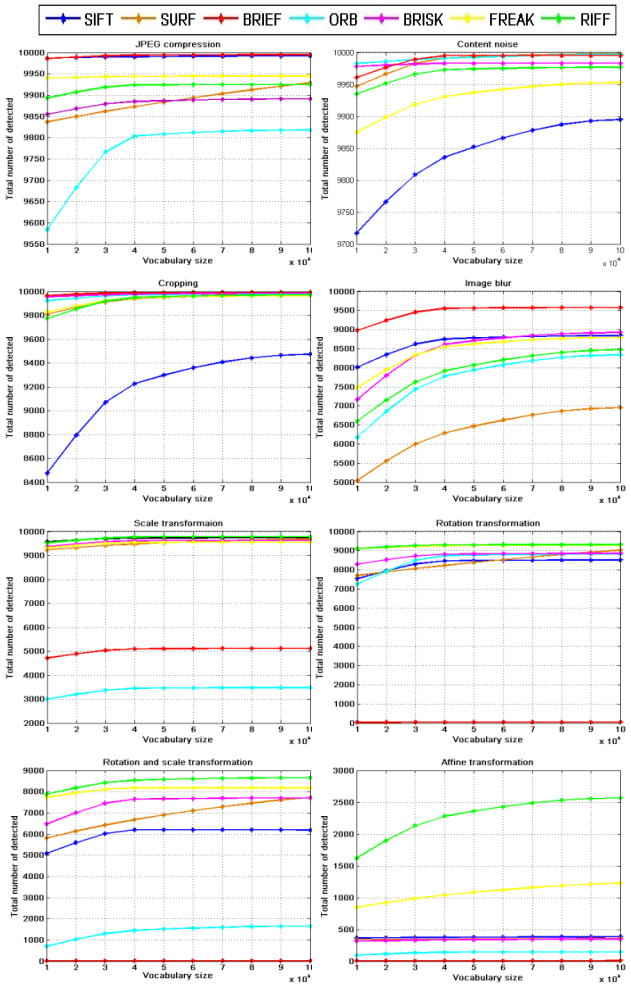


Figure 7. Number of detected duplicates in each transformation category.

5. CONCLUSIONS

We have proposed a novel salient point descriptor named RIFF which was inspired by the sampling pattern from the human eye (We make no claims of biological relevance). The main contribution of the RIFF descriptor is in constructing the descriptor so that the discriminatory power is optimized by ranking and deleting points with low distinctiveness. From our bag of words image retrieval tests on three well known datasets, RIFF outperformed the other feature descriptors with respect to scale, rotation, and affine transformations.

6. REFERENCES

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

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

[3] 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.

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

[5] 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.

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

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

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

[9] Tola, E., Lepetit, V., and Fua, P. 2010. Daisy: an Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), 32(5):815–830, 2010.

[10] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., and Zisserman, A. 2010. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascal-network.org/ challenges/VOC/voc2008

[11] Griffin, G., Holub, A., and Perona, P. 2007. The Caltech 256. California Institute of Technology, Technical Report.

[12] Huiskes, M.J., Thomee, B., and Lew, M.S. 2010. New Trends and Ideas in Visual Concept Detection. *ACM International Conference on MIR*, 527-536.

[13] O'Hara, S., and Draper, B. A. 2013. Are you using the right approximate nearest neighbor algorithm?. *IEEE Workshop on Applications of Computer Vision (WACV)*, 9-14.

[14] Salton G., and McGill, M.J. 1986. Introduction to Modern Information Retrieval. McGraw-Hill.