Large scale visual search

Wu, S.

Chapter 3

RIFF: Retina-inspired Invariant Fast Feature Descriptor

In this chapter, we first present the Retina-inspired Invariant Fast Feature, RIFF, which is designed for invariance under scaling, rotation, and affine image deformations. The RIFF descriptor is based on the comparison of the intensity of pair-wise pixels over a sampling pattern that has similarities with the human retina. Then we introduce a strategy to improve 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 benchmark datasets demonstrates that the RIFF descriptor has competitive performance compared to state-of-the-art descriptors. Additionally, a popular approach from literature is to use visual words (or Bag-of-Words) constructed from real valued local descriptor (SIFT and SURF). To accommodate large scale data sets, we used an approximate nearest neighbor (ANN) based clustering approach to both real valued local descriptors and binary string local descriptors (BRIEF, ORB, BRISK, FREAK, Binboost and LATCH). The results on these test sets reveal that some of the recent binary string approaches outperform notable descriptors such as SIFT and SURF.

3.1 Introduction

Efficiently establishing the correspondences between images is very useful for numerous applications of computer vision, such as content-based image search, image classification, object tracking, and panorama stitching. Salient point methods are leading approaches, which have 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 (e.g., blob, region, or point) in images which typically are in some way informative. The descriptor gives a model or representation of a local image region. Prior research of salient points has focused on high repeatability detectors and robustness under scaling and rotation [108].

The SIFT descriptor [14] is the most popular salient point method. It computes the Difference-of-Gaussian (DoG) operator in the Gaussian scale space, and assigns an orientation and a descriptor to each salient point based on the local gradient histogram. The SURF [12] salient point detector makes use of a box-filter to achieve efficient extrema detection in the scale space and it performs well with respect to the criterion 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 that have specific advantages such as low memory requirements as well as computationally efficient matching using the Hamming distance (bitwise XOR followed by a bit count). BRIEF [117] first uses Gaussian smoothing on the selected image patch, and creates a binary string descriptor by the comparison of the intensities of randomly sampled pixel-pairs around the patch center. ORB [118] employs the most efficient FAST [96] detector to determine the salient points in different layers of an image pyramid. It use the intensity centroid algorithm to determine the orientation for each point. The binary string descriptor of ORB is determined similar to BRIEF and effectively improves the robustness under image rotation and scale changes. BRISK [16] applies a 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 [17] 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. Moreover, some local binary descriptors based on a supervised learning scheme also show good performance (BinBoost [123] and LATCH [124]).

The recently introduced salient point descriptors each have specific strengths. Some are most restrict to scale changes, whereas others are designed for speed and/or low memory requirements. Our goal was to design a descriptor which was optimized to be robust under affine image transformations including rotation and scaling. In this chapter, we first propose a novel discriminative salient point descriptor which is named "RIFF" because the sampling pattern is inspired by the distribution of cones (color vision) that can be observed in the human eye.

Moreover, empirical experiments conducted over the past decade have demonstrated that one of the most popular and successful approaches towards image similarity and visual concept analysis is to use salient point algorithms combined with visual word model and an approximate nearest neighbors (ANN) search [122]. This is mainly due to the robustness of salient point descriptors under various geometric transformations and to the introduction of the visual word model, which significantly improved the search efficiency and the adaptability to a particular image dataset. Current visual words systems are predominantly built using salient points algorithms such as SIFT and SURF whose descriptors are real valued. In contrast to the real valued descriptors, binary string descriptors were proposed in order to generate and use the feature descriptors in a more efficient way (e.g., BRIEF, ORB, BRISK, FREAK, BinBoost and LATCH). Another goal of this chapter is to give insights into the performance and requirements of these descriptors for large scale image search.

The main contributions of this chapter are as follows:

First, we proposed a salient point descriptor which outperforms current methods regarding robustness under affine image 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 a large amount of descriptors and increasing accuracy.

Second, we compared several of the most promising local descriptors on a wide variety of near duplicate transformations within the visual words paradigm. This is very important for computer vision applications because each social application may involve a different set of image transformations. Our results give some insight into which descriptors would be better or worse candidates in each of these cases. To our knowledge, this is the first contribution that compares visual word models generated by recent binary string features and applies on large scale image copy detection.

Third, we made a comparison of different types of features in terms of feature extraction and vocabulary generation by measuring, for example, computational efficiency as well as memory efficiency. This requirements are important because in some situations speed might be more important than accuracy alone. In addition we adopted the ANN search to achieve the vocabulary generation.

The rest of the chapter is organized as follows: In Section 3.2, we present the generation of our RIFF local feature descriptor. In Section 3.3, we describe the details of the visual word model generation. The datasets and evaluation criteria in the experiment are described in Section 3.4. The performance results of the proposed descriptor compared to current state-of-the-art descriptors are shown in Section 3.5, and finally conclusions are given in Section 3.6.

3.2 Discriminate RIFF Local Descriptor

3.2.1 Retina Sampling Pattern Review

The retina sampling pattern is based on the topology of the human retina as found in neuroscience research. This research reveals that the spatial distribution density of cone cells in the retina decreases exponentially with the increasing

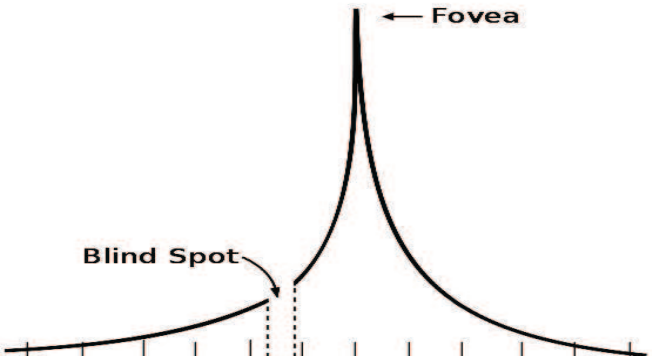


Figure 3.1: Illustration of the density distribution of cones in the human retina.

distance to the center of the fovea. Moreover, it is believed that the image signals pass through from cone cells to ganglion cells, where the receptive field of each ganglion cell uses the Difference of Gaussian (DoG) model with various sizes and that encodes differences into action potentials. Our approach employed a similar retina sampling pattern, which places different sizes of blocks at the defined locations in the pattern. The illustration of the cones density can be seen in Figure 3.1.

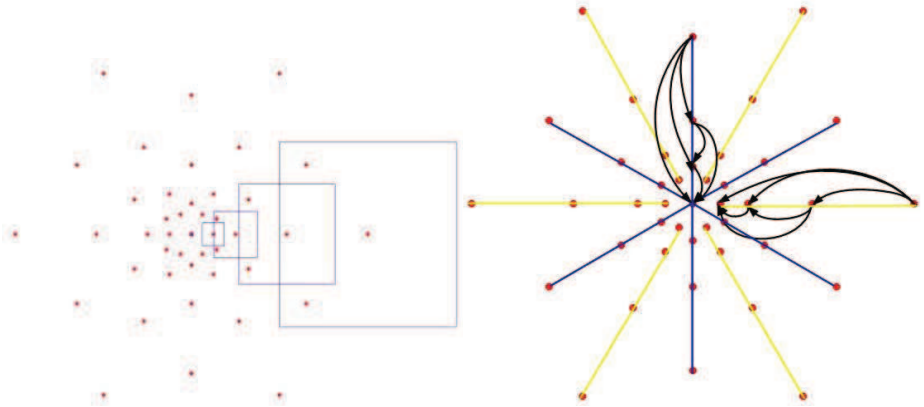
Inspired by recent work that use decreasing circular polar densities in diverse applications ranging from stereo matching to object recognition [16, 17, 125], the sampling pattern for RIFF in 2D decreases exponentially as shown in Figure 3.2 (a).

3.2.2 Descriptor Generation

3.2.2.1 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



(a) (b)

Figure 3.2: (a) The 2D exponentials decreasing polar sampling pattern for RIFF with N=43 points: the red points denote the sampling point locations, the blue rectangle represents a receptive field, and the size of the rectangle corresponds to its Gaussian kernel which is 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.

geometry which estimates the orientation using local gradients: $\triangle y$ and $\triangle x$ and then determine the angle from the arctangent of $(\triangle y/\triangle x)$ (for details see FREAK [17]). We also estimate the local gradients by pair-wise differences between equidistant points from the center of the retina sampling pattern.

3.2.2.2 Descriptor Generation

The procedure of RIFF descriptor generation is different from previous salient point approaches such as BRIEF, ORB, BRISK and FREAK, which compare the pixel-pair intensities 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, where v_i is a real 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 results in the dimension of the descriptor d equal to 72. For clarity, we have

Figure 3.3: Matches (blue lines), between images after an affine viewpoint change, found by using the SIFT (OpenCV) salient point approach.

displayed in Figure 3.2 (b) 1 of the 6 pair-wise comparisons on the blue axes and 1 of the 6 pair-wise on the yellow axes where each black curve denotes one pair-wise comparison. Since we place a block at each sampling point, the integral image (summed area tables) was used for computational efficiency. It was not necessary for RIFF to compare the intensity of all possible $N \times (N-1)/2$ sampling pairs, which was necessary when calculating the binary string features used in previous methods. Moreover, the dimension of RIFF is smaller than SIFT, which may improve the speed of indexing and matching.

3.2.2.3 Discriminative Strategy

Even though location, scale and orientation have been estimated, current salient point detectors have difficulty with affine viewpoint changes such as depicted in Figure 3.3. 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.

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

K nearest neighbors in the image.

$$D_p(p \in I) = \sum_{j=1}^{K} \|f_p - f_i\|_2 \tag{3.1}$$

$\|.\|_2$ denotes the Euclidean distance. Intuitively, a higher discriminatory score demonstrates that the feature of point p is more distinctive than features of near by other points. The parameter K is set to 2 in the experiment, as it can achieve a good performance at a very low computational complexity. Furthermore, we use an exponential function in order to emphasize the discriminative score:

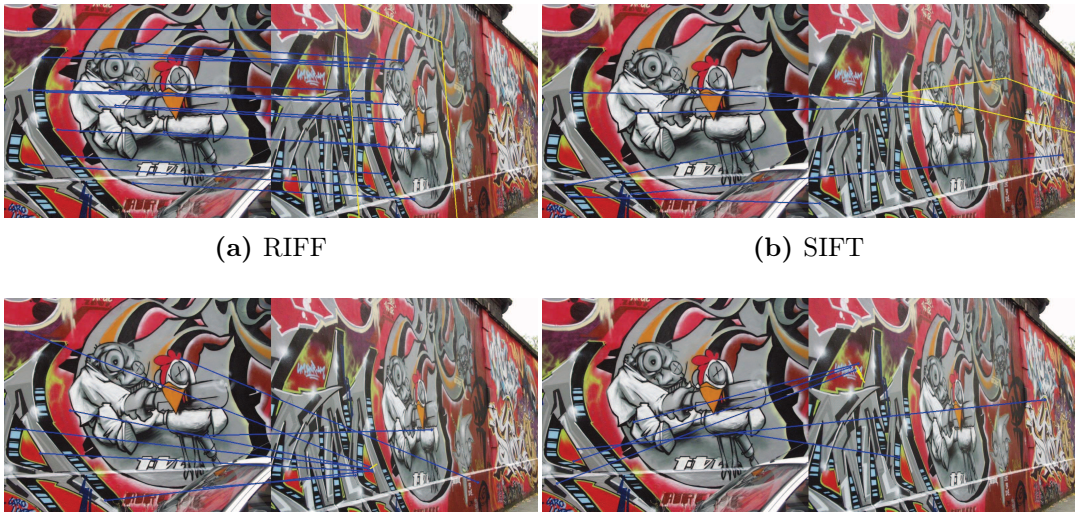
$$D_p^{'}(p \in I) = \exp(-\lambda \cdot |D_p|) \tag{3.2}$$

$|\cdot|$ denotes the normalization of D_p (in the range $[0, 1]$), λ is a weight of discriminative score and set to 6 that can achieve a good performance in the experiment. We note that after the above process, a smaller $D_p^{'}$ score correlates to more distinctive feature points, so we can sort these scores and define a threshold to filter out unstable salient points. The final set is a smaller number of discriminative features which are more robust to various image transformations, while reducing required subsequent processing, e.g., descriptor indexing as well as dictionary learning in large scale image applications.

We set the value of threshold in the *NNDR* to 0.75, and the homography between two compared images is estimated by the RANSAC algorithm. In preliminary tests, RIFF exhibited competitive performance for image copies detection under affine image transformations in comparison to the popular SIFT, SURF, and recent FREAK descriptors as shown in Figure 3.4.

3.3 Visual Word Model based Image Search

There are billions of images available on the WWW, scientific databases and private collections that do not have sufficient annotations for broad and accurate searching. Moreover, the number of images is ever increasing, and a large number



(a) RIFF **(b)** SIFT

(c) SURF **(d)** FREAK

Figure 3.4: Illustration of descriptor matching. Here RIFF is compared to SIFT, SURF, and FREAK on an image from a challenge on affine object detection (Graffiti 1-5 proposed by Mikolajczyk and Schmid [108]).

of similar copies exist. These copies can be viewed as transformed versions of the original images. Since common transformations such as geometric distortions, compression, crop, and color space changes could easily result in numerous copies or near-duplicates, it is a major challenge to achieve accurate, time and space efficient large scale detection of duplicates. Conventional global feature based image representations (color histogram, textual feature and shape information) can be used to perform an image search. However, they can not handle complex image transformations, such as rotation and scale changes. The visual word model based image representation (BoW [126], Fisher Vector [2] and VLAD [127]) takes advantage of the high discriminative capability of local descriptors in different contents and the applicability of different similarity measures to address complex image changes.

Visual word models, inspired by the field of information retrieval, were established by the introduction of salient point local descriptors, mainly because those local descriptors were shown to be invariant to scaling, rotation and noise. A visual word model represents an image as a histogram of visual words through feature

quantization and significantly improves the accuracy of image retrieval and object classification.

Typical implementations [4, 128, 129, 130] of the visual word model start by detecting salient points or regions from all images in the dataset and generate a descriptor for each salient point or region. These descriptors can be further clustered into a vocabulary consisting of visual words where each cluster center represents a visual word, and the size of the vocabulary is equal to the number of clusters. Based on salient points extracted as salient image patches, an image is frequently represented using a histogram according to the occurrence frequency of each visual word.

For the popular real valued local descriptors (e.g., SIFT, SURF), the simple K-means clustering algorithm can be used to train the visual word vocabulary. The initialization of cluster centers is first generated by randomly choosing candidates from the descriptors group. After that, at the beginning of each iteration, the remaining descriptors are assigned to their closest cluster center. The center can be updated by the mean value of the assigned descriptors. Euclidean distance is used as a distance measure in the assignment procedures.

For binary string descriptors, the Hamming distance metric is used. As it only use bitwise XOR followed by a bit count, it offers a higher matching speed. As the traditional computation of an average is not suitable for binary features, we employed an approach named "K-majority" [131] to calculate the mean value of binary string descriptors.

The K-majority method refines cluster centers based on the statistics of the total number of 1's at the same bit position among all the descriptors belonging to the same cluster. Suppose a cluster consist of I binary string features: $F_i, i = 1, 2, ..., I$, and we treat a binary feature as $F = [bit_1, bit_2, ..., bit_J]$, $1 \leq j \leq J$, where J denotes the length of binary string feature. The following function can then be used to update the cluster center.

$$score(bit_j) = \sum_{i=1}^{I} F_i(bit_j)/I \tag{3.3}$$

$$Center(bit_j) = \begin{cases} 1, & if : score(bit_j) \geq 0.5 \\ 0, & if : score(bit_j) < 0.5 \end{cases} \tag{3.4}$$

Function (3.4) implies that if the number of 1's is larger than half the number of total descriptors belonging to the specified cluster, the new value of the same bit position of the center is set to "1", otherwise it is set to "0".

However, it is a challenge to apply the flat K-means or flat K-majority to large scale vocabulary construction, because it is computationally expensive to perform clustering in high dimensional spaces. In order to reduce the computational complexity of linear search, an approximate nearest neighbors approach (ANN) was adopted to assign the labels of optimal cluster centers to descriptors. Compared with the flat K-means and flat K-majority, ANN-based K-means and K-majority approaches could effectively reduce the complexity from $O(NK)$ to $O(Nlog(K))$ during each iteration, where N denotes the number of descriptors, and K is equal to the number of centers. Considering the different properties of real valued descriptors and binary string descriptors, ANN search is based on a KD-tree index and a LSH index respectively [132]. The LSH index space is based on multi-probe LSH, which has the advantage of reduced storage requirements. Once the visual vocabulary has been obtained, we represent an image as a bag of visual words according to the popular *tf-idf* weighting scheme [133]. The *tf-idf* weighting scheme can reduce the contributions of common visual words, while at the same time increasing the contributions of discriminative words. Through building an index for the image features in the dataset, a ranked list of search results could be efficiently returned according to the distance similarity with the query image feature.

3.4 Experimental Results

The experimental environment for the evaluation is an Intel Quad Core i7 Processor (2.67GHz), 12GB of RAM, 64-bit OS. The implementations of BinBoost is from the author, others are implementations from OpenCV. The parameters of

each salient point method were set to the defaults. We used 8 randomized forests in the KD-tree index, 20 hash tables in the multi-probe LSH index. Our evaluation implementations are available at: http://press.liacs.nl/researchdownloads/.

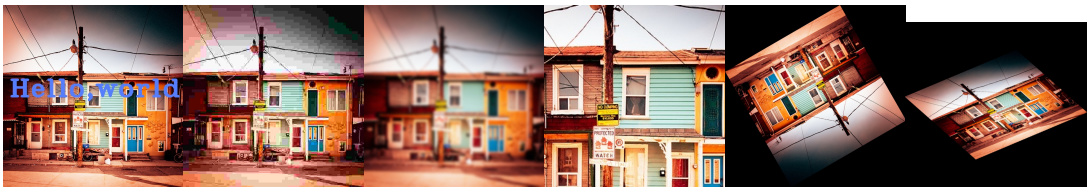
3.4.1 Datasets and Evaluation Criteria

The evaluation of visual word based large scale image copy detection is performed on three image datasets: PASCAL VOC2012 [134], Caltech 256 [135], and MIR FLICKER 1Million [136]. Moreover, a series of near duplicates were created for the test. We use 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, rotation, scale and affine deformation: rotation + scale + 3D perspective distortion.

Scale change: we resized the original images by changing the scale factors from 20% to 200% with a step size of 20%.
Cropping: starting with a 50×50 pixel central region in the image, the width and height of the cropped area of the image is gradually increased by 10 pixels.
Image compression: JPEG compression copies are produced by setting the image quality factors in the range from 95% to 5%.
Text noise: images are modified by adding various sizes and colors of text in the central area.
Image blur: A series of blurred images is created by smoothing the image using Gaussian smoothing.
Deformation: includes several subsets where image copies (rotation, rotation together with scale, and viewpoint transformation) are created by rotation as well as perspective distortion with different angles.

A total number of 1000 images in each dataset are randomly selected as query images, and 80 duplicates of each query image are generated. Some examples of each dataset used for evaluation are illustrated in Figure 3.5.

(a)



(b)

Figure 3.5: Examples from each dataset for the evaluation of salient point methods. (a) Examples from the VOC, Caltech 256 and MIR datasets. (b) Examples of generated duplicates: text noise, JPEG compression, image blur, image crop, rotation and affine transformation, respectively.

3.4.2 Evaluation of Image Copy Detection

In this section, we focus on constructing the visual word vocabulary not only by using real valued descriptors, but also binary string descriptors. We use ANN search to efficiently train the vocabulary. We first compared the proposed RIFF with a number of the most promising salient feature descriptors on a wide variety of near duplicate transformations within the visual words paradigm. This is the most important part of this section because each application may involve different image transformations and our results give some insight into which descriptors would be better or worse candidates. Then, we made a comparison of different types of features in terms of feature extraction and vocabulary generation by measuring indicators of computational efficiency as well as space requirements. There characteristics are valuable because in some situations speed, for example, might be more important than accuracy alone.

In order to make an objective comparison of different types of local descriptors, we also choose to use the same detector for each local descriptor. SURF was applied as the salient point detector, and we combine the SURF detector with various fea-

ture descriptors including SIFT, SURF, ORB, BRIEF, BRISK, FREAK, RIFF, BinBoost and LATCH in the evaluation. The performance of various vocabularies is evaluated in terms of computational efficiency, memory requirements, and accuracy.

The criterion to estimate the similarity of two images represented by visual words is via the cosine distance measure. We use mAP (mean Average Precision) as a criterion to evaluate the performance of search accuracy.

3.4.2.1 Evaluation of Time and Storage

We first focus on the efficiency and space requirements of generating the vocabularies for the different descriptor types. For this evaluation we use the PASCAL VOC dataset. Ten million salient point features were extracted from the dataset.

Figure 3.6 illustrates the computational efficiency of different types of vocabulary generation as well as the storage space requirement under different cluster sizes. The vocabulary generation based on the compared descriptors all reveal an almost linear growth with increasing vocabulary size. In Figure 3.6 we can see that the execution time of the vocabulary training stage with real valued descriptors is nearly 4 times faster than that of binary string features, however, binary type vocabularies have significantly lower space requirements.

3.4.2.2 Evaluation of Search Accuracy

We evaluated the performance of image copy detection using various visual vocabularies. As all the generated duplicates are added into the datasets, the scale of PASCAL and Caltech is roughly 10 thousand, and MIRFLICKR contains around one million. The comparison experiment with different types of vocabulary is based on varying the vocabulary size.

Overall, the RIFF based visual words model outperformed the other descriptors on the PASCAL VOC, Caltech 256 and MIRFLICKR-1M datasets as shown in

Figure 3.7. The mAP score results also demonstrate that binary local descriptor based visual vocabularies offer good performance. Comparing the evaluation results on the one million-scale dataset and the results on VOC dataset, there is no significant mAP score decrease when the data size increased from ten thousand to more than one million. We can also note that the FREAK descriptor based vocabulary has better mAP score on average across the three datasets than other binary string based vocabularies. Below we will discuss how various descriptor based visual word models performed under the different transformations.

Our goal of this part is to determine the robustness of the visual vocabularies to various image transformations. As shown in Figure 3.8, RIFF had the best performance on the distortions related to scale, rotation, and affine transformations. It showed average performance on blurring and showed competitive performance on the rest of the transformations. When the transformation keeps the structure in place such as blur and JPEG compression, SIFT has high accuracy but was outperformed by BRIEF, while BinBoost showed a weakness for the cases of blur and JPEG compression. We observe that when pictorial information is added to or deleted from an image copy, SIFT was consistently outperformed by the other descriptors. Specifically, FREAK performed well on transformations which deformed the image structure such as affine transformations or combining rotation with scaling. BRIEF showed particularly poor performance on rota-

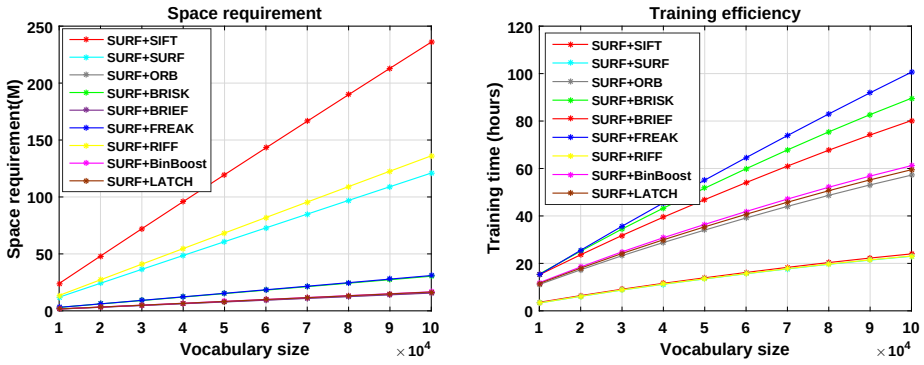


Figure 3.6: Comparison of different descriptors in terms of time efficiency and space requirements during the training. Both space requirement and training time show almost linear growth when the size of the vocabulary increases.

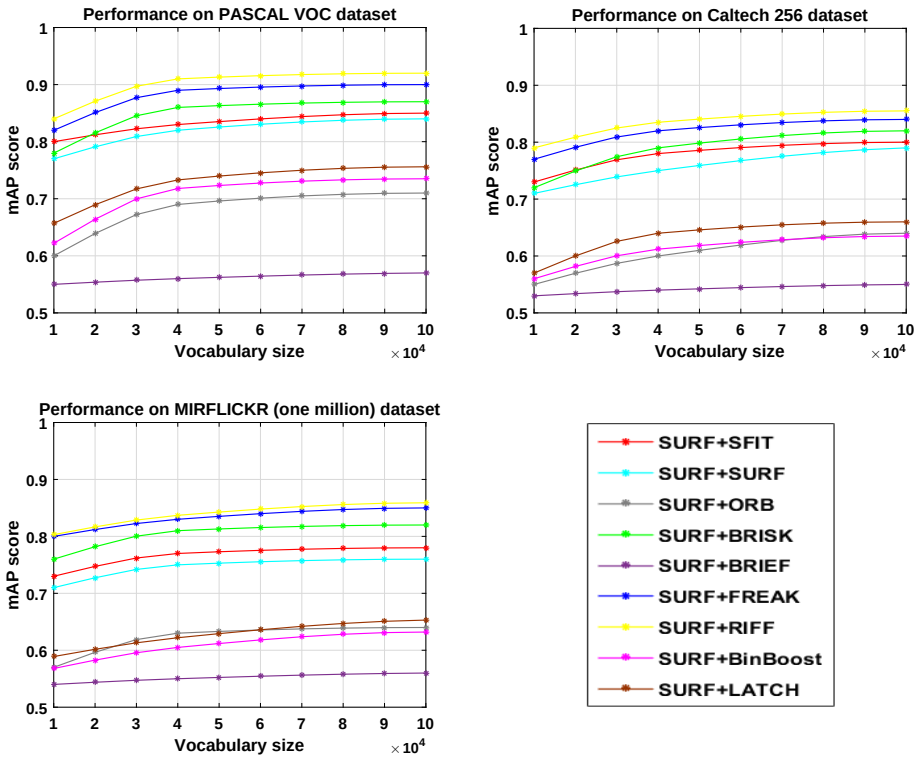


Figure 3.7: Detection accuracy (mAP score) on three datasets (PASCAL VOC, Caltech 256 and MIRFLICKR). The size of the vocabulary varies from 10000 to 100000.

tion transformations and LATCH showed a poor performance on scale changes. Note that the affine deformation represents the most difficult category, as the total number of detected copies is extremely low for all types of compared visual vocabularies.

According to the copy detection accuracy, the robustness of visual word model based image representations mainly rely on the capability of the local descriptor. We can see that the BRIEF descriptor is not rotation and scale invariant, thus, a visual word model trained on BRIEF is sensitive to rotation and scale changes. The ORB descriptor makes an improvement in case of the rotation changes when compared to the BRIEF descriptor, therefor, vocabularies trained on the ORB descriptors showed better performance than BRIEF. RIFF, BRISK and FREAK based visual word models have high performance for rotation and scale invari-

ance, probably because the local descriptors of RIFF, BRISK and FREAK use circular sampling patterns. The vocabularies trained on new binary string features (BRISK and FREAK) and the real valued features (SIFT, SURF and RIFF) all are scale invariant and robust to JPEG compression and blur noise. For the category of learning based local descriptors, the learning scheme of BinBoost is not robust to the JPEG compression and blur noise. LATCH does not use scale information during the learning process.

3.5 Conclusions

We have proposed a novel salient point descriptor named RIFF which was inspired by the sampling pattern used by 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. Our Bag-of-Words image retrieval tests on three well known datasets, showed RIFF outperforming the other feature descriptors with respect to robustness to scale, rotation, and affine transformations. Furthermore, we presented a performance evaluation of real valued and binary string salient point descriptors. The time complexity and space requirements showed that binary string descriptors are efficient in terms of feature extraction time and memory usage. Regarding the criterion of the mAP score, the image copy detection experiments showed some significant strength of binary string local descriptors. FREAK clearly outperformed SIFT on rotation and scale, and affine transformations. BRIEF had the best accuracy in case of image blur and was among the best in case of image cropping.

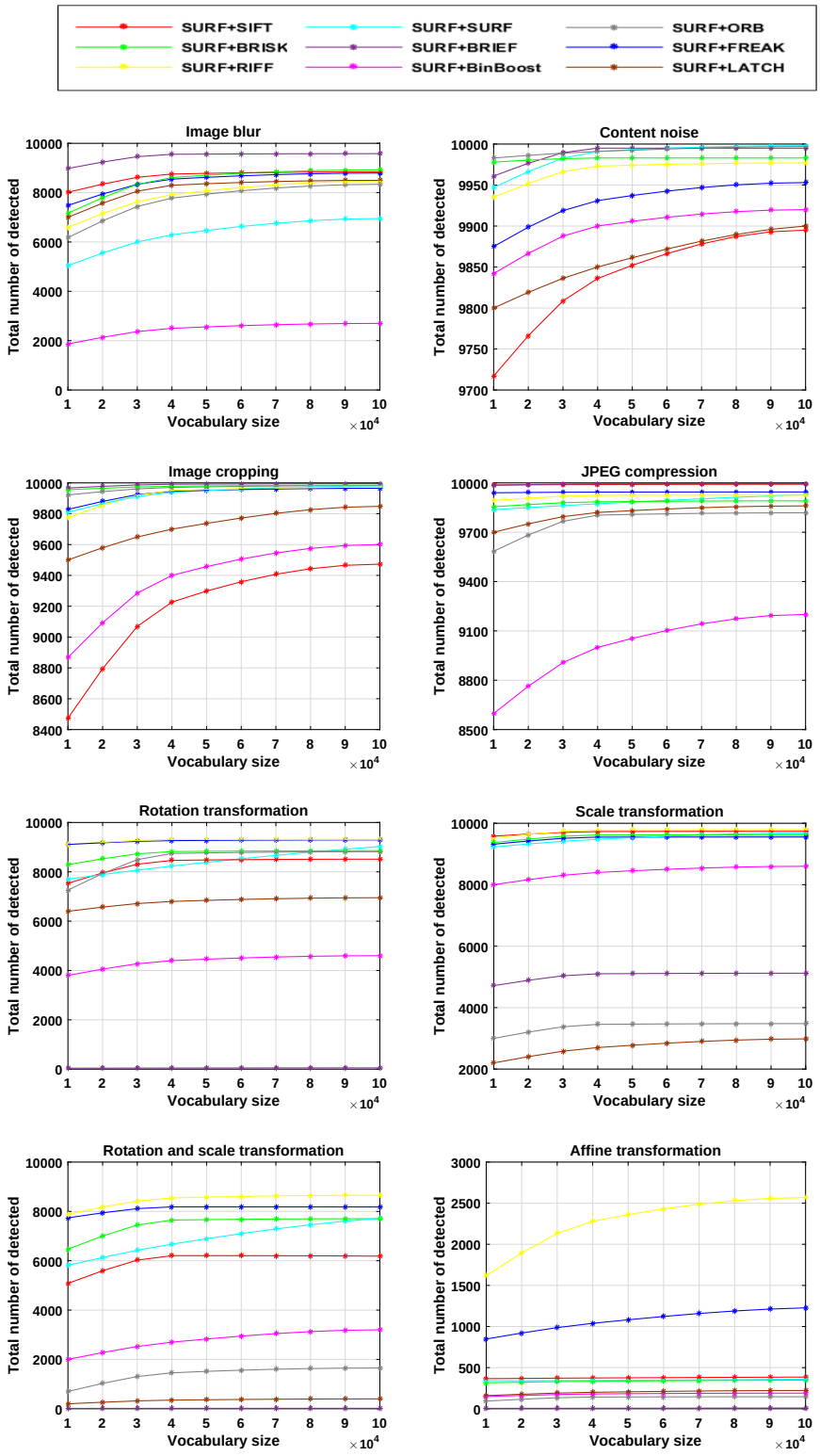


Figure 3.8: Total number of detected duplicates from different types and different sizes based vocabularies in each transformation category. The size of vocabulary varies from 10000 to 100000.