ON AGGREGATION OF LOCAL BINARY DESCRIPTORS

Syed Husain and Miroslaw Bober

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK.
sh0057, m.bober@surrey.ac.uk

ABSTRACT

This paper addresses the problem of aggregating local binary descriptors for large scale image retrieval in mobile scenarios. Binary descriptors are becoming increasingly popular, especially in mobile applications, as they deliver high matching speed, have a small memory footprint and are fast to extract. However, little research has been done on how to efficiently aggregate binary descriptors. Direct application of methods developed for conventional descriptors, such as SIFT, results in unsatisfactory performance. In this paper we introduce and evaluate several algorithms to compress high-dimensional binary local descriptors, for efficient retrieval in large databases. In addition, we propose a robust global image representation; Binary Robust Visual Descriptor (B-RVD), with rank-based multi-assignment of local descriptors and direction-based aggregation, achieved by the use of L1-norm on residual vectors. The performance of the B-RVD is further improved by balancing the variances of residual vector directions in order to maximize the discriminatory power of the aggregated vectors. Standard datasets and measures have been used for evaluation showing significant improvement of around 4% mean Average Precision as compared to the state-of-the-art.

Index Terms— visual search, binary descriptors, global descriptor, image retrieval

1. INTRODUCTION

Many contemporary pipelines for object recognition and retrieval choose to employ local binary descriptors in order to reduce extraction and matching complexity. This is particularly important in mobile applications, where at least some processing is performed on a terminal with limited resources; but even when server resources are available computational complexity is still an issue due to the ever increasing scale of databases, image resolutions and the required accuracy and speed of search. Consequently, local binary descriptors become increasingly popular, as they deliver high matching speed, small memory footprint and are relatively fast to extract. However, descriptor-by-descriptor matching and matching based on inverted index is too complex for mobile visual search (MVS). Hence the majority of MVS systems have to rely on global descriptors. While many techniques exist for extracting global representations from floating-point local descriptors, such as SIFT [1], surprisingly hardly any research exists on how to efficiently aggregate local binary descriptors. Binary descriptors, such as ORB [2], FREAK [3] and BRISK [4], are significantly faster to compute compared to SIFT and even SURF [5], and deliver comparable performance. However direct application of aggregation methods developed for conventional floating-point descriptors results in suboptimal or even poor performance.

In this paper we conduct an in-depth evaluation of applicability of various existing aggregation schemes to binary descriptors and propose a novel aggregation technique, which delivers state-of-the art performance. Our scheme builds on the original VLAD [6] descriptor, and its extension by Husain and Bober [7] which combined VLAD aggregation with rank-based multi-assignment and robust norm on residual vectors. Here we further extend their approach by introducing cluster-level de-correlation and whitening of normalized residual vectors. We show that the proposed extension brings significant improvement to performance on all datasets, for all operating points and input binary feature combination.

The key contributions of this paper include (i) analysis of the behavior and performance of existing aggregation schemes with local binary features, (ii) a new B-RVDW pipeline (shown in Figure 1) delivering superior performance and (iii) its optimization with intensity-based (BRISK, FREAK) and gradient-based (BRIGHT) binary descriptors.

The remainder of this paper is organized as follows. In Section 2, we briefly review existing techniques for aggregation of local descriptors and discuss previous work on binary descriptors. We also outline three designs of local descriptors that we subsequently evaluate in our pipeline. In Section 3 we introduce our global descriptor based on binary local features, which will be referred to as B-RVD and propose a novel extension by whitening normalized residual vectors forming B-RVDW descriptor. Results of an extensive evaluation are presented in Section 4. Finally, conclusions and future work are summarized in Section 5.

This research has received partial funding from the European Commission 7th Framework Programme under grant agreement Nr. 610691.

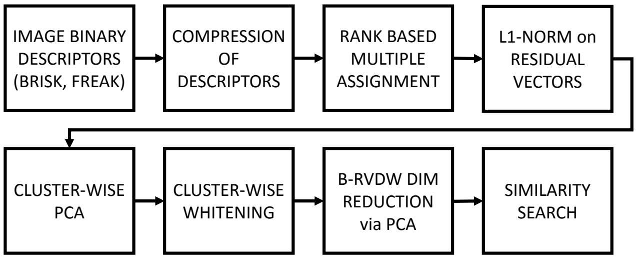


IMAGE BINARY DESCRIPTORS (BRISK, FREAK)
COMPRESSION OF DESCRIPTORS
RANK BASED MULTIPLE ASSIGNMENT
L1-NORM on RESIDUAL VECTORS
CLUSTER-WISE PCA
CLUSTER-WISE WHITENING
B-RVDW DIM REDUCTION via PCA
SIMILARITY SEARCH

Fig. 1. B-RVDW extraction pipeline using rank-based multi-assignment, residual normalization, cluster-wise whitening and B-RVDW dimensionality reduction

2. GLOBAL REPRESENTATIONS FROM LOCAL BINARY FEATURES

The Bag-of-Words (BoW) [8] is frequently used to create a fixed-length global image representation. It is a histogram, where each local descriptor is assigned to the nearest visual word. The histogram is normalized using tf-idf weighting and retrieval can be performed using an inverted list. Peronnin et al. [9] employed Fisher kernels to aggregate local image descriptors into compact vector representations, called Fisher Vectors (FV). This assumes that samples of local descriptors are distributed according to the Gaussian Mixture Model (GMM). Jegou et al. [6] proposed a non-probabilistic version of FV called VLAD, that builds an image representation by aggregating residual vectors for descriptors, clustered according to proximity criterion in the feature space. All the above representations were originally derived for the SIFT descriptor, which is a 128-dimensional floating-point vector formed by computing local image gradient histograms. We pose here an important question - how will these representations perform when derived directly from binary features? Can they match the performance of global descriptors based on floating-point features? What is the best pipeline and the optimal parameters? These are non-trivial questions, as binary descriptors exhibit very different characteristics and many underlying assumptions are clearly no longer fulfilled. For example, it is known that while the centering operations and least squares criteria of PCA are well suited to real-valued data (such as SIFT), they are not generally appropriate for binary data types [10].

There are few works addressing aggregation of binary descriptors. Grana et al. [11] integrated an ORB descriptor into the BoW model for image classification. Opdenbosch et al. [12] adapted VLAD to the 256D BRIEF binary descriptor [13]. While their approach works with a limited database in a navigation context, they unfortunately do not present any evaluation on reference databases.

For our investigation, we select three local binary descriptors: two intensity-based (BRISK [4] and FREAK [3]) and one gradient-based (BRIGHT [14]). Our choice is motivated by their high level of performance in retrieval using descriptor-by-descriptor matching with bi-directional ratio test, but without geometric verification. We also want to verify if the type of features used (intensity-test vs histogram of gradients) impacts performance. In all cases we employ BRISK key-point detection, as it is fast, delivers good performance and is the de-facto standard in mobile applications. The binary descriptors used in this study are briefly summarized below.

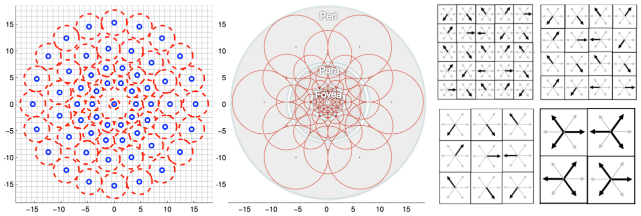


Fig. 2. Sampling patterns for BRISK (left), FREAK (mid) and BRIGHT (right).

The binary BRISK descriptor is constructed by concatenating the results of simple brightness comparison tests. The test pattern, shown in Fig 2-left, defines L locations equally spaced on four concentric circles centered at the key-point. Gaussian smoothing with standard deviation σ_i proportional to the distance between the points on the corresponding circle i is used to increase robustness by reducing aliasing. FREAK also uses a circular sampling grid, however the density of points is not constant (Fig 2-center) - it drops exponentially when moving away from the center, as in the human retina. The difference with BRISK is the exponential change in sampling density and the overlap in receptive fields. The 512 pairs are selected using an algorithm similar to ORB to be un-correlated and hence highly discriminant. The BRIGHT descriptor is based on a hierarchical Histogram of Gradients (HOG). Three layers of hierarchical pyramid of HOG, with different block partitioning of 5×5, 4×4, and 3×3 are extracted (Fig 2-right). The elements of the histogram are binarized, and a subset of bits may be progressively selected to form a scalable descriptor with a size between 32-150 bits. We used the full 150 bit descriptor in our study. Compared to BRISK and FREAK, the BRIGHT descriptor is three times more compact and exhibits a similar level of performance in descriptor-by-descriptor matching tests; it is also similar to SIFT as it uses histograms of gradients. The next section presents the design of our aggregation scheme and demonstrates its performance with the three selected binary features.

3. AGGREGATION OF BINARY DESCRIPTORS

Before designing the aggregation pipeline, experiments are performed to investigate the type of dimensionality reduction that should be applied to the input binary vectors. The experiments have shown that direct aggregation in the binary domain (i.e keeping the input vectors and cluster centers in

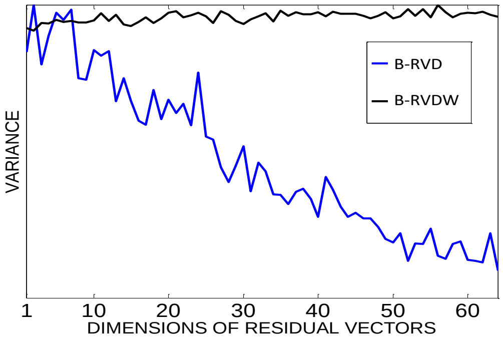


Fig. 3. Energy in each dimension of weighted residual vectors z_{tj} before aggregation into a global descriptor.

binary format) delivers generally poor results. This is particularly prominent when the dimensionality of the input descriptors is significantly reduced, e.g. from 512 to 128 or 64 for the BRISK or FREAK. The best results are achieved by applying PCA directly to the binary vectors and using PCA to reduce dimensionality. Dimensionality reduction has a very significant impact, beyond what can be explained as selection of the dominant components.

Based on the conclusions from the initial experiments, we design our pipeline, which is presented in Figure 1. More precisely, let $\mathcal{X} = \{x_t \in \mathbb{R}^d, t = 1...T\}$ be the set of binary local descriptors, such as BRISK or FREAK, extracted from an image. The descriptors are compressed to d' dimensions using bit selection or Principal Component Analysis (refer Section 4). The compressed descriptors are rank-assigned to multiple clusters and a robust representation of residual vectors in each cluster is derived forming the B-RVD global descriptor. The high-dimensional global descriptor is converted into a compact signature by application of PCA.

In the B-RVD aggregation scheme each binary descriptor x_t is defined by its position with respect to the K nearest cluster centers (typically K=3) in the high dimensional space. More precisely, K-means clustering is performed to learn a codebook of $\{v_1, ..., v_n\}$ of n cluster centers typically between 64 and 512. Each local descriptor x_t is quantized to K nearest cluster centers thus increasing the number of descriptors assigned to each center, resulting in more populous cluster-level representations, which are more robust. For each cluster, the residual vectors $x_t - v_j$ are computed and subsequently L1-normalized. The application of L1-norm on residual vectors ensures that the influence of each local descriptor on the cluster-level representation is comparable. The normalized residual vectors are weighted for each neighborhood rank (N) before aggregation to reflect the fact that the descriptors belonging to rank-1 nearest neighbor (N1) are considered more reliable and stable - the reliability and stability of features decreases as we increase the neighborhood rank. The neighborhood weights w_N are computed as the empirical probability that two descriptors forming a matching pair (inliers) with specific neighborhood rank are assigned to the same cluster. In the B-RVD representation, the weights are: $1, 0.5$ and 0.25 for the assignments with rank one, two and three respectively. The weighted residual vectors z_{tj} are computed as:

$$z_{tj} = w_N \frac{x_t - v_j}{||x_t - v_j||_1} \tag{1}$$

The cluster level representation δ_j is computed by aggregating vectors z_{tj} across all N. Each δ_j is L2-normalized [15] and concatenated to form the final B-RVD descriptor R. The dimensionality of B-RVD is $D = d' \times n$.

$$\delta_j = \sum_{N=1}^{K} \sum_{x_t \in rankNof\ v_j} w_N \frac{x_t - v_j}{||x_t - v_j||_1} \tag{2}$$

$$R = \left\{ \frac{\delta_1}{||\delta_1||_2}; \frac{\delta_2}{||\delta_2||_2}; ...; \frac{\delta_n}{||\delta_n||_2} \right\} \tag{3}$$

Binary RVD Local Whitening (B-RVDW)

The variance in each dimension of vector z_{tj} is different which affects the discriminability of the B-RVD representation. We solve this problem by applying cluster level PCA and a whitening operation on z_{tj} vectors before aggregation. Given a set of m weighted residual vectors $z_{1j}, z_{2j}, ..., z_{mj}$ in $\mathbb{R}^{d'}$ extracted from training images I, we form the column data matrix $Z_j = (z_{1j}, z_{2j}, ..., z_{mj})$ for each cluster j.

1) *Centering data:* To center the data in matrix Z_j, the first step is to compute the mean vector $\mu_j = \mathbb{E}[z_{tj}]$.

$$\mu_j = \frac{1}{m} \sum_{I} \sum_{N=1}^{K} \sum_{x_t \in rankNof v_j} w_N \frac{x_t - v_j}{||x_t - v_j||_1} \tag{4}$$

Every z_{tj} vector is subtracted by the mean vector μ_j, i.e., $z_{tj} \leftarrow (z_{tj} - \mu_j)$, then we get the centered column data matrix $Z_j^c = (z_{1j}, z_{2j}, ..., z_{mj})$.

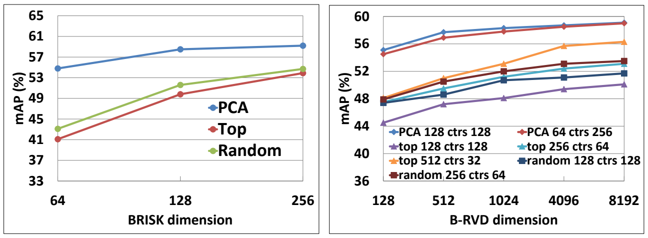
2) *PCA projection matrix:* Based on the centered vectors $z_{1j}, z_{2j}, ..., z_{mj}$, the covariance matrix is computed:

$$\Sigma_j = \frac{1}{m} \sum_{t=1}^{m} z_{tj} z_{tj}^{\top} = \frac{1}{m} Z_j^c Z_j^{c\top} \tag{5}$$

For each cluster j, we compute a PCA matrix P_j whose columns consists of the eigenvectors of Σ_j corresponding to the d' largest eigenvalues $\lambda_1 \geq \lambda_2 ... \geq \lambda_{d'}$. Finally, the cluster level whitening matrix P_j^w is computed as, P_j^w=$P_j {\Omega_j}^{-\frac{1}{2}}$, where $\Omega_j = diag(\lambda_1, \lambda_2 ..., \lambda_{d'})$.

3) *B-RVDW representation:* Given an image I, the vectors z_{tj} are extracted for each cluster j. The mean subtracted z_{tj} vector is projected using P_j and subsequently whitened before aggregation into δ_j.

$$\delta_j = \sum_{N=1}^{K} \sum_{x_t \in rankNof\ v_j} P_j^{w\top} \left(w_N \frac{x_t - v_j}{||x_t - v_j||_1} - \mu_j \right) \tag{6}$$



(a) (b)

Fig. 4. (a) Compression of Brisk descriptor. (b) Impact of binary descriptor compression and vocabulary size.

The L2-normalized δ_j vectors are stacked to form the final B-RVDW representation R^w. Figure 3 shows the energy distribution in each dimension of residual vectors before aggregation using B-RVD and B-RVDW representations. In B-RVD the energy in each dimension of z_{tj} is different while in B-RVDW, after performing PCA+Whitening the energy is balanced between dimensions leading to improved performance.

Compact Global Descriptor

The global descriptor can be compacted by performing dimensionality reduction via PCA while retaining its discriminative power. The compact B-RVDW vector R^s is computed using equation $R^s = P^\top \times (R^w - R_0)$, where R_0 is the mean of the signatures of R^w and P is $D \times D'$ matrix ($D' \leq D$) of eigenvectors associated with the largest eigenvalues of the covariance matrix of signatures of R^w. The similarity between R^s vectors is computed using standard Euclidean distance.

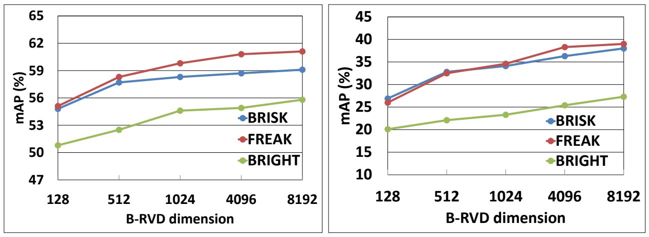
4. EXPERIMENTAL EVALUATION

The purpose of this section is to establish B-RVD and B-RVDW as the state-of-the-art global image representations. First we present the datasets and evaluation measures used for benchmarking. Then we investigate the impact of dimensionality reduction techniques and the optimal size of vocabulary. Finally we perform retrieval experiments to compare the performance of several global image representations including VLAD+, FV, B-RVD and B-RVDW.

Datasets. The performance is evaluated on two standard image retrieval benchmarks: the INRIA Holidays and the Oxford building dataset. Independent datasets were used for all learning stages, to eliminate any bias.

The INRIA Holidays dataset [16] contains 1491 high resolution holiday photos with 500 of them used as queries. Retrieval accuracy is measured by mean Average Precision (mAP), as defined in [17].

The Oxford5k dataset [17] contains 5062 images gathered from Flickr by searching for particular Oxford landmarks. This dataset has been manually annotated to generate 55 query images corresponding to 11 different landmarks. The performance is evaluated using mAP.



(a) (b)

Fig. 5. Comparison of BRISK, FREAK and BRIGHT descriptors (a) Holidays dataset (b) Oxford dataset.

Unless otherwise stated, key-points were detected using a multi-scale version of the AGAST detector [4]. This detector is fast but does not compensate for affine distortions, hence our results cannot be directly compared to the results obtained with Hessian-affine detector which helps to rectify the orientation ambiguity of affine-covariant points[18].

Impact of descriptor compression and vocabulary size

Aggregating 512-dimensional binary descriptors (e.g. BRISK or FREAK) using a small vocabulary of $n = 64$ visual words results in a 32k-dimensional global descriptor. This size is prohibitive due to memory and complexity constraints, hence it is necessary to reduce dimensionality of the input descriptors. We investigate the optimum parameters using the BRISK descriptor and the following three compression approaches: (i) PCA, (ii) Random bit selection, (iii) Top bit selection. Figure 4 (a) shows the performance of the B-RVD representation on the Holidays dataset as a function of the dimension after compression (d'). It can be observed that projecting the binary descriptor using PCA provides significantly better performance compared to selecting dimensions. There is only a relatively small gain in performance (+0.5%) by increasing the output PCA dimensions from 128 to 256.

The full optimization of the B-RVD pipeline requires exhaustive experiments of all different combinations of parameters. To reduce the complexity of the analysis and keep the experimental load reasonable we keep the dimensionality of B-RVD representation fixed at 16k and search for the best combination of input descriptor dimensionality and number of clusters n. Figure 4 (b) shows the performance of B-RVD on the Holidays dataset using different combinations of d' and n where (PCA 128 ctrs 128) indicates that the dimensionality of a binary descriptor is reduced from 512 to 128 via PCA and aggregation is performed using 128 clusters. It can be observed from that the best performance is achieved at **(PCA 128 ctrs 128)**, the same operating point is also optimal for the FREAK descriptor.

Brisk vs Freak vs Bright

Figure 5 shows the results obtained on the Holidays and Oxford datasets with B-RVD vector generated using BRISK, FREAK and BRIGHT descriptors. The input descriptor di-

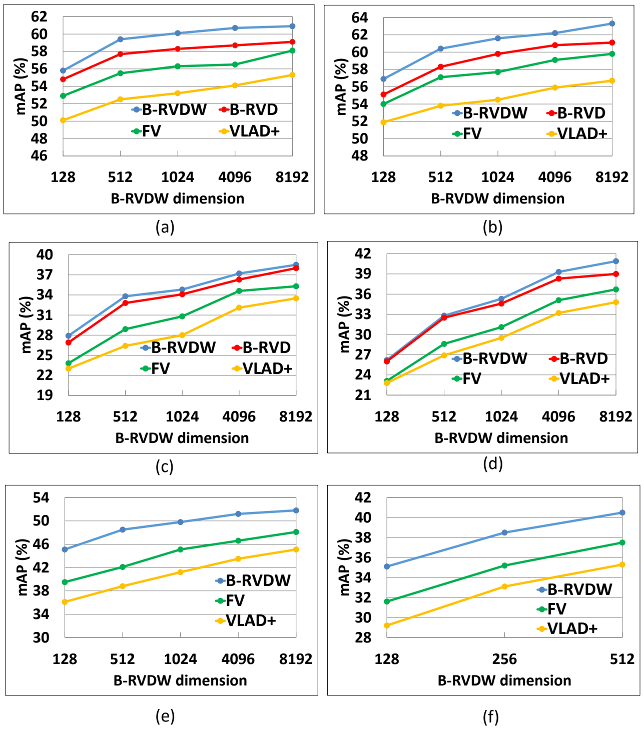


Fig. 6. Performance of global descriptors on Holidays dataset using (a) BRISK (b) FREAK . Performance of global descriptors on Oxford dataset using (c) BRISK (d) FREAK (e) Holidays100k using FREAK and (f) large scale experiments on Holidays1Million using FREAK.

mensionality and visual vocabulary is fixed at $d' = 128$ and $n = 128$ respectively. It can be seen that the combination FREAK+B-RVD outperforms BRISK+B-RVD for high-dimensional global descriptors (D'=8192, 4096 and 1024). The performance is similar when the B-RVD dimension is reduced to 512 and 128. The performance of BRIGHT is significantly worse than BRISK and FREAK. This is most likely due to the fact that the original dimensionality of BRIGHT, at 150 bits, is too low to be discriminative.

Evaluation of global descriptors

We compare our best representation B-RVDW with B-RVD, FV and VLAD+. The VLAD+ [19] representation is a modified version of VLAD in which each residual vector is L2-normalized before aggregation. The global descriptor dimensionality is $128 \times 128 = 16384$. It can be observed from Figure 6 (a-d) that B-RVDW significantly outperforms all representations using both FREAK and BRISK as input descriptors. Compared to FV, B-RVDW offers an average gain of +3.4% and +3.8% in mAP on the Holidays and Oxford datasets. We also compared the performance of FREAK+B-RVDW and FREAK+FV using large scale dataset of Holidays + Flickr 1M distractors. Figure 6 (f) clearly shows that B-RVDW significantly outperforms FV also on large scale.

Method	Det.	Desc.	Dim	Oxf	Hol
FV	BRISK	BRISK	8k	35.3	58.1
FV	BRISK	FREAK	8k	36.7	59.8
VLAD+	BRISK	BRISK	8k	33.5	55.3
VLAD+	BRISK	FREAK	8k	34.8	56.7
B-RVD	BRISK	BRISK	8k	38.1	59.1
B-RVD	BRISK	FREAK	8k	39.0	61.1
B-RVDW	BRISK	BRISK	8k	38.5	60.9
B-RVDW	BRISK	FREAK	8k	**40.9**	**63.3**
BoW [6]	HA	SIFT	20k	35.4	43.7
FV [6]	HA	SIFT	8k	41.8	60.5
VLAD [6]	HA	SIFT	8k	37.8	55.6
VLAD [20]	DoG	SIFT	8k	24.3	56.1

Table 1. Comparison of the B-RVDW with the state-of-the-art (Det:Detector, Desc:Descriptor, Dim:Dimensions)

Detector	Descr.	Loc. desc. extr. (ms)	Global desc. encoding (ms)	Total time
BRISK	BRISK	85	200	285
BRISK	FREAK	85	200	285
DoG	SIFT	900	190	1090
HA	SIFT	1230	190	1420

Table 2. Average time required to compute B-RVDW signature using different detectors+descriptors combinations (DoG: Difference of Gaussian, HA: Hessian-Affine)

The upper section of Table 1 lists the performance of binary descriptor aggregation schemes with the fast BRISK detector and BRISK/FREAK descriptors. It can be seen that B-RVDW significantly outperforms the state of the art global descriptors by $+3.5\%$ on average.

Although this paper is about aggregation of binary descriptors, we also compare our framework with global descriptors that use Hessian-affine or DoG detector with SIFT descriptor, which is fives times slower (lower part). It can be clearly observed that the FREAK+B-RVDW is better or comparable to HA+SIFT and DoG+SIFT combined with BoW, VLAD and FV. On large-scale dataset of Holidays1Million, FREAK+B-RVDW achieves **35.1%** compared to SIFT+FV 31.8% [6].

Complexity Analysis

Table 2 compares the average time required to compute B-RVDW signature, for different combinations of detectors and descriptors. The total time comprises local descriptors extraction (first column) and encoding of the global representation (second column). It can be observed that use of binary local descriptors reduced computational complexity by factor of 5, as compared to working with SIFT.

5. CONCLUSIONS

We presented an in-depth investigation of aggregation strategies for local binary descriptors, based on three high-performance descriptors: BRISK, FREAK and BRIGHT, in combination with the BRISK key-point detector. We have compared various strategies to dimensionality reduction and demonstrated the utility of the front-end PCA, which converts binary input vectors into intermediate floating-point representations. Furthermore, we proposed a new binary aggregation pipeline, extending the B-RVD representation by introducing whitening of the normalized residual vectors, which leads to significant gains in performance compared to the state-of-the-art FV and VLAD. The new pipeline was optimized and evaluated, demonstrating improved mAP of 4% in large scale retrieval, and significantly reduced extraction complexity.

6. REFERENCES

[1] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[2] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *International Conference on Computer Vision*, 2011.

[3] Raphael Ortiz, "Freak: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[4] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *International Conference on Computer Vision*, 2011.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, 2008.

[6] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1704–1716, 2012.

[7] S. Husain and M. Bober, "Robust and scalable aggregation of local features for ultra large-scale retrieval," in *IEEE International Conference on Image Processing*, 2014.

[8] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.

[9] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors.," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] A. Schein, L. Saul, and L. Ungar, "A generalized linear model for principal component analysis of binary data," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[11] Costantino Grana, Daniele Borghesani, Marco Manfredi, and Rita Cucchiara, "A fast approach for integrating orb descriptors in the bag of words model," *Proc. SPIE*, 2013.

[12] D. Van Opdenbosch, G. Schroth, R. Huitl, S. Hilsenbeck, A. Garcea, and E.G. Steinbach, "Camera-based indoor positioning using scalable streaming of compressed binary image signatures.," in *IEEE International Conference on Image Processing*, 2014.

[13] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "Brief: Binary robust independent elementary features," in *European Conference on Computer Vision*, 2010.

[14] K. Iwamoto, R. Mase, and T. Nomura, "Bright: A scalable and compact binary descriptor for low-latency and high accuracy object identification," in *IEEE International Conference on Image Processing*, 2013.

[15] R. Arandjelović and A. Zisserman, "All about VLAD," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[16] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, 2010.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[18] M. Perďoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[19] J. Delhumeau, P. Gosselin, H. Jégou, and P. Pérez, "Revisiting the VLAD image representation," in *ACM Multimedia*, 2013.

[20] E. Spyromitros-Xioufis, S. Papadopoulos, Kompatsiaris I., G. Tsoumakas, and I. Vlahavas, "A comprehensive study over vlad and product quantization in large-scale image retrieval," *IEEE Transactions on Multimedia*, 2014.