Improving large-scale image retrieval through robust aggregation of local descriptors

Syed Husain and Miroslaw Bober, *Member, IEEE*

Abstract—Visual search and image retrieval underpin numerous applications, however the task is still challenging predominantly due to the variability of object appearance and ever increasing size of the databases, often exceeding billions of images. Prior art methods rely on aggregation of local scale-invariant descriptors, such as SIFT, via mechanisms including Bag of Visual Words (BoW), Vector of Locally Aggregated Descriptors (VLAD) and Fisher Vectors (FV). However, their performance is still short of what is required. This paper presents a novel method for deriving a compact and distinctive representation of image content called Robust Visual Descriptor with Whitening (RVD-W). It significantly advances the state of the art and delivers world-class performance. In our approach local descriptors are rank-assigned to multiple clusters. Residual vectors are then computed in each cluster, normalized using a direction-preserving normalization function and aggregated based on the neighborhood rank. Importantly, the residual vectors are de-correlated and whitened in each cluster before aggregation, leading to a balanced energy distribution in each dimension and significantly improved performance. We also propose a new post-PCA normalization approach which improves separability between the matching and non-matching global descriptors. This new normalization benefits not only our RVD-W descriptor but also improves existing approaches based on FV and VLAD aggregation. Furthermore, we show that the aggregation framework developed using hand-crafted SIFT features also performs exceptionally well with Convolutional Neural Network (CNN) based features. The RVD-W pipeline outperforms state-of-the-art global descriptors on both the Holidays and Oxford datasets. On the large scale datasets, Holidays1M and Oxford1M, SIFT-based RVD-W representation obtains a mAP of 45.1% and 35.1%, while CNN-based RVD-W achieve a mAP of 63.5% and 44.8%, all yielding superior performance to the state-of-the-art.

Index Terms—visual search, image retrieval, local descriptor aggregation, global descriptor

1 Introduction

The explosive growth in the multimedia industry has created a need for effective and computationally efficient retrieval systems. Given a large collection of images and videos, the aim is to retrieve individual images and video shots depicting instances of a user-specified object (query). There are a range of important applications for image retrieval including management of multimedia content, mobile commerce, surveillance, augmented automotive navigation etc. Despite formidable efforts, the performance of existing systems still lacks in terms of robustness, processing speed and detection rates; especially at the low false alarm rates required for immense databases.

The task of performing robust, accurate and scalable visual search is challenging. An object's appearance can depend on many compounding factors such as object scale, illumination, occlusions, dissimilar backgrounds, varying viewpoints and compression artifacts. Additionally, today's systems must be highly scalable due to the huge volumes of multimedia data, which can comprise billions of images.

A classical approach to object based image retrieval involves use of scale-invariant local descriptors such as SIFT [1] or later variants [2], [3], which achieve some robustness to scale changes, illumination conditions and occlusions. While the use of local descriptors increases the robustness against large visual distortions and partial occlusions, it also increases computational complexity of the search, as they need to be individually compared and matched. One solution is to form a single, global image representation, thereby simplifying the matching process and leading to an improved matching speed and lower memory usage.

Such global representation can be extracted either directly from pixels in an image or by aggregating local image descriptors. The first category, represented for example by the GIST descriptor [4], has a significant drawback in that it lacks robustness to many common image transformations, such as partial occlusion, cropping and rotation [5]. In contrast, successful techniques for image retrieval tend to focus on deriving image representations from local descriptors. For example, the Bag-of-Words (BoW) representation has been widely used in textual document classification and subsequently adapted to computer vision tasks, including retrieval [6]. Perronnin et al. [7] proposed a local descriptor aggregation based on the Fisher kernel framework, while Jegou et al. introduced a simplified version called VLAD [8]. Section 2 reviews global descriptors in detail.

In this paper we propose a novel aggregation scheme; Robust Visual Descriptor, with various extensions. The core RVD concept originates from robust statistics, and was introduced in a preliminary way in [9], [10], [11]. Here we present in-depth details of the core method, significantly expand the experimental evaluation, and - crucially - reveal new insights on the reasons underpinning its strong performance. We then further extend the core method by introducing cluster-wise whitening and novel descriptor normalization, leading to world-class performance, significantly out-performing any results published to date. Our main contributions include:

- *M.Bober and S.Husain are with the Centre for Vision, Speech and Signal Processing (CVSSP), Department of Electrical Engineering, University of Surrey, Guildford, Surrey, GU2 7XH, UK.*
  E-mail: {sh0057, m.bober}@surrey.ac.uk

- The core RVD aggregation approach employing a novel rank-based multi-assignment with a direction-based aggregation method.
- An RVD extension, where the variances of residual vector directions are balanced, in order to maximize the discriminatory power of the aggregated vectors. This is achieved by a novel intra-stage pre-processing of the residual directions using cluster-wise PCA with a whitening operation. We call this representation RVD-W.
- A new normalization approach applied after the RVD-W vectors are transformed via global PCA. Our normalization involves L1-norm followed by a power-norm. We show that the aforementioned normalization is different from existing approaches including Whitening with L2 normalization [12] and power+L2 normalization [13], and offers significant benefits in terms of retrieval accuracy.
- We conduct an in-depth experimental study to illustrate the effects of various elements of the RVD-W pipeline, in order to understand the roots of its superior performance. For instance, we analyze the behavior of the rank-based multi-assignment and compare it to the well studied hard-assignment of VLAD and the soft-assignment used in the FV approach. We also investigate the impact of applying L1-normalization to residual vectors, cluster-wise PCA and cluster-wise whitening before aggregation into RVD-W.
- We combine our RVD-W framework with CNN-based deep features, demonstrating performance beyond the state of the art [14] [29] [34]. In particular, we show that RVD-W aggregation outperforms both FV and sum-pooling methods for CNN features, contrary to the recent views that sum-pooling [14] is preferred with deep features. Uniquely, we also present CNN based results on datasets with 1M distractors.

This paper is organized as follows: Section 2 provides insights into state-of-the-art local descriptor aggregation methods. In Section 3, we present our core RVD representation and its variants. The experimental setup and the detailed evaluation of our method is presented in Section 4. In Section 5 we compare our results with the state of the art demonstrating significant improvement over recent global descriptors on both the Holidays and Oxford datasets. On the large scale datasets, Holidays1M and Oxford1M, SIFT-based RVD-W representation obtains a mAP of 45.1% and 35.1%, while CNN-based RVD-W achieve a mAP of 63.5% and 44.8%, all significantly outperforming any results published up to date.

2 Global Descriptors

This section reviews state-of-the-art global representations that encode the distribution of local image descriptors in an image, namely the BoW, Fisher Vector, VLAD and Triangulation Embedding.

2.1 Bag of Words (BoW)

The Bag of Words representation is essentially a histogram, where each local descriptor is assigned to the nearest cluster or visual word. In the training stage, which is performed off-line, a codebook $\{\mu_1, ..., \mu_n\}$ of n cluster centers is learned via K-means clustering. Given an image, the extracted descriptors are vector quantized into a predefined visual vocabulary. To form a fixed length n-dimensional representation of an image, a histogram of local descriptors with n bins (visual words) is constructed, where each descriptor is assigned to the closest (in the Euclidean space) cluster. The inverse document frequency (idf) [15] weighting is typically applied and an inverted list is used for efficient comparison of BoW representations. Several advances have been made to improve BoW scalability and robustness. Approximate K-means clustering [16] algorithm was proposed to produce large and discriminative vocabularies. The robustness of the BoW system was improved by using a soft assignment technique [17].

2.2 Fisher Vectors (FV)

Fisher Vector encoding aggregates local image descriptors based on the Fisher Kernel framework. More precisely, let $\mathcal{X} = \{x_t \in \mathbb{R}^d, t = 1...T\}$ be the set of local descriptors, such as SIFT [1] or CNNs [18], extracted from an image I. Let u_Θ be an image-independent probability density function which models the generative process of $\mathcal{X}$, where Θ represents the parameters of u_Θ.

Fisher Vector framework assumes u_Θ to be a Gaussian Mixture Model (GMM) [7]: $u_\Theta(x) = \sum_{j=1}^{n} \omega_j u_j(x)$. We represent the parameters of the n-component GMM by $\Theta = (\omega_j, \mu_j, \Sigma_j : j = 1..n)$, where $\omega_j, \mu_j, \Sigma_j$ are respectively the weight, mean vector and covariance matrix of Gaussian j. The covariance matrix of each GMM component j is assumed to be diagonal and is denoted by σ_j^2. The GMM assigns each descriptor x_t to Gaussian j with the soft assignment weight (τ_{tj}) given by the posteriori probability:

$$\tau_{tj} = \frac{exp(-\frac{1}{2}(x_t - \mu_j)^T \Sigma_j^{-1}(x_t - \mu_j))}{\sum_{i=1}^{n} exp(-\frac{1}{2}(x_t - \mu_i)^T \Sigma_i^{-1}(x_t - \mu_i))} \tag{1}$$

The GMM can be interpreted as a probabilistic visual vocabulary, where each Gaussian forms a visual word or cluster. The d-dimensional derivative with respect to the mean μ_j of Gaussian j is denoted by ζ_j:

$$\zeta_j = \frac{1}{T\sqrt{\omega}_j} \sum_{t=1}^{T} \tau_{tj} \frac{x_t - \mu_j}{\sigma_j} \tag{2}$$

The FV representation ζ_Θ of image I is obtained by concatenating the gradients ζ_j for all Gaussian $j = 1..n$ and is therefore D=$d \times n$ dimensional.

Compared to the BoW, which only records the count of local descriptors in each visual word, the FV encodes the higher order statistics, resulting in a more discriminative representation and hence better performance.

While the FV descriptor is considered to represent the state-of-the art, it has its own limitations and can be improved upon significantly. We believe that these limitations arise from several factors. Firstly, local descriptors actually rarely follow GMM distribution in the feature space, impacting negatively on the model performance. Secondly, as we show later in this paper, local descriptor cluster assignments in FV often degrade to a single-assignment, reducing the overall robustness to noise and outliers. This problem is addressed by our ranked-based multi-assignment.

2.3 Vector of Locally Aggregated Descriptors (VLAD)

Jegou et al. [8] proposed a simplified version of FV called VLAD. A codebook $\{\mu_1, ..., \mu_n\}$ of n cluster centers is obtained via K-means clustering and each descriptor $x_t \in \mathbb{R}^d$ is hard-assigned to its nearest cluster center $NN(x_t)$. The main idea here is to compute the cluster level representations $\zeta_j \in \mathbb{R}^d$ by aggregating the differences $x_t - \mu_j$ (i.e. residual vectors) between the descriptors and their corresponding cluster centers:

$$\zeta_j = \sum_{x_t : NN(x_t) = j} x_t - \mu_j \tag{3}$$

The D-dimensional VLAD is obtained by concatenating all aggregated vectors ζ_j for all clusters $j = 1, .., n$.

The main drawback of VLAD is its limited robustness to outliers. A single outlier descriptor located far from the cluster center can outweigh the combined contribution from many inlier descriptors located close to that center.

Recently several improvements have been made to the original VLAD representation. Husain et al. [10], [11] introduced the RVD global descriptor with rank based multiple assignment and L1-norm on residual vectors leading to significant gain in retrieval performance. In [19], Arandjelovic et al. introduced: (i) intra-normalization where the sum of residual vectors within a cluster is L2-normalized, (ii) extraction of multiple VLAD descriptors from sub-regions in an image, and (iii) a vocabulary adaptation algorithm to correct inconsistent visual vocabularies. Delhumeau et al. [20] proposed to rotate the L2-normalized residual vectors inside each cluster center according to the local PCA basis. In [21], Xioufis et al. proposed to aggregate SURF and color SURF (CSURF) descriptors into a VLAD vector and observed that a VLAD+SURF significantly outperforms a VLAD+SIFT pipeline. Eggert et al. [22] suggested to apply cluster-wise PCA on aggregated residual vectors before concatenation and named their representation HVLAD. In [23], Liu et al. proposed the Hierarchical VLAD (HiVLAD), where the cluster centers obtained from the K-means clustering are further divided into sub-clusters and a VLAD vector is computed for each sub-cluster. Picard et al. [24], introduced Vector of Locally Aggregated Tensors (VLAT) descriptor, formed by aggregating tensor products of local descriptors. In [25], Negrel et al. proposed two extensions to VLAT: (i) PCA cluster-wise VLAT (PVLAT), which applies PCA to each flattened cluster representation and (ii) Compression of the PVLAT vector (CPVLAT).

2.4 Triangulation Embedding (TEmb) and Function Approximation embedding (FAemb)

Recently, Jegou et al. [13] proposed a local descriptor aggregation scheme using triangulation embedding. In this approach, each local descriptor x_t is hard-assigned to all cluster centers. The residual vectors $x_t - \mu_j$ are computed and subsequently L2-normalized to yield a set s_{tj} of normalized residual vectors:

$$s_{tj} = \left\{ \frac{x_t - \mu_j}{||x_t - \mu_j||_2} \right\} \quad \text{for} \quad j = 1...n, \tag{4}$$

where $||.||_2$ denote L2-norm. The vectors in set s_{tj} are stacked to form representations $S_t = [s_{t1}^\top, ...s_{tn}^\top]^\top$ and each S_t is whitened (centered, rotated and scaled based on eigenvalues) to form triangulation embedding $\phi_{\Delta t} \in \mathbb{R}^{nd}$:

$$\phi_{\Delta t} = \Sigma^{-1/2}(S_t - S_0) \tag{5}$$

where S_0 and Σ are respectively the expected value and covariance matrix associated with S. The $\phi_{\Delta t}$ vectors are aggregated using sum aggregation to form a global image representation ψ_s:

$$\psi_s(\mathcal{X}) = \sum_{x_t \in \mathcal{X}} \phi_{\Delta t} = \Sigma^{-1/2} \left(\sum_{x_t \in \mathcal{X}} S_t \right) - T\Sigma^{-1/2} S_0 \tag{6}$$

One drawback of sum aggregation is that the vector ψ_s is more influenced by common uninformative local descriptors rather than rare but informative ones. This problem is alleviated by aggregating descriptors $\phi_{\Delta t}$ using democratic aggregation where a weight φ_t is applied to each $\phi_{\Delta t}$ before aggregation into global signature ψ_d. The weights φ_t are learned using the modified Sinkhorn algorithm:

$$\psi_d(\mathcal{X}) = \sum_{x_t \in \mathcal{X}} \varphi_t \phi_{\Delta t} \tag{7}$$

In [26], Do et al. introduced a local descriptors embedding approach named Function Approximation embedding (FAemb). In this method, each descriptor x_t is assigned to a set of n anchor points $C = \{\mu_1, ..., \mu_n\}$. The tensor products $(x_t - \mu_j)(x_t - \mu_j)^\top$ are computed and the upper triangles (including the diagonals) are unfolded to yield a set s_{tj} of vectors:

$$s_{tj} = \{\delta_{\mu_j}(x_t) V((x_t - \mu_j)(x_t - \mu_j))^\top\} \quad \text{for} \quad j = 1...n, \tag{8}$$

where $\delta_{\mu_j}(x_t)$ is coefficient corresponding to μ_j of descriptor x_t and $V(H)$ is a function that flattens the matrix H to a vector. As in TEmb, the vectors in the set s_{tj} are concatenated to form S_t and each S_t is whitened to yield FAemb representation $\phi_{Ft} \in \mathbb{R}^{nd(d+1)/2}$. An image signature is computed by aggregating ϕ_{Ft} vectors using democratic aggregation.

Experimental results presented in [13] show that Triangulation Embedding performs very well on all standard benchmarks, with $\phi_\Delta + \psi_d$ considerably outperforming $\phi_\Delta + \psi_s$. However, while the complexity of $\phi_\Delta + \psi_s$ is similar to FV, $\phi_\Delta + \psi_d$ is typically two orders of magnitude slower due to the complexity of Sinkhorn algorithm. The overall computational cost is prohibitively high; in fact it prevented the authors from performing experiments on datasets larger than 100K using 64 cluster centers. Also, as we show later in the experimental Section 4, $\phi_\Delta + \psi_d$ performance deteriorates rapidly when the dataset size increases beyond 1M, particularly when the descriptor dimensionality is reduced, making it unsuitable for large scale retrieval. The computational complexity of FAemb signature $\phi_F + \psi_d$ approach is even higher than $\phi_\Delta + \psi_d$; no large scale results are reported in [26].

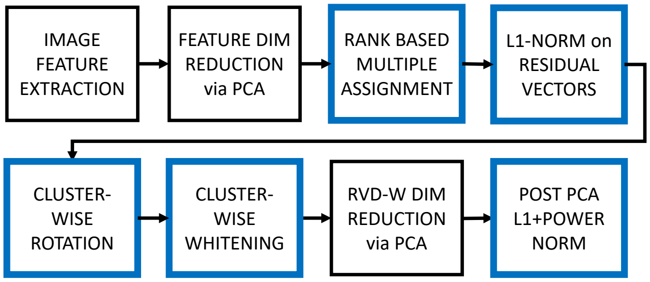


IMAGE FEATURE EXTRACTION
FEATURE DIM REDUCTION via PCA
RANK BASED MULTIPLE ASSIGNMENT
L1-NORM on RESIDUAL VECTORS
CLUSTER-WISE ROTATION
CLUSTER-WISE WHITENING
RVD-W DIM REDUCTION via PCA
POST PCA L1+POWER NORM

Fig. 1. RVD-W extraction pipeline using rank-based multi-assignment, residual normalization, cluster-wise whitening and post PCA processing

3 Robust Visual Descriptor (RVD)

Our RVD representation is inspired by concepts from Robust Statistics. In retrieval, image pairs with matching visual objects contain a certain proportion of matching local descriptors, contaminated by a large proportion of non-matching outliers. For example, in the Oxford dataset, the median percentage of inliers is only 20%. Thus the task of image matching may be considered as detection of matching local descriptor pairs in the strong presence of outliers. The aim is therefore to develop a global representation of a set of local descriptors that will be both representative and robust in the mathematical sense, i.e. not affected by a large number of additional local descriptors.

The core RVD, introduced in a preliminary fashion in [9], [10], [11] builds a global image representation by aggregating the normalized residual vectors for descriptors that are rank-assigned to multiple cluster centers. This manuscript provides in-depth technical details, formulation and evaluation of the core RVD including a new study explaining why rank-based assignment is so effective. Furthermore, we significantly improve the performance of the core RVD representation by balancing the energy of the weighted residual vector dimensions. This is achieved by de-correlating and subsequently whitening the weighted residual vectors before aggregation into a cluster-wise RVD-W representation. In addition, the RVD-W global descriptor is projected via PCA and post-processed by L1-norm and power-norm, to increase the separability between matching and non-matching global descriptors. In order to reduce the memory requirement and complexity, the RVD-W signature is encoded using the Optimized Product Quantization (OPQ) approach [27]. Finally, we also demonstrate how to effectively aggregate CNN-based features into the RVD-W representation. We will now describe in detail the components of the RVD-W pipeline, as shown in Figure 1.

Local descriptor assignment

RVD is a global image representation formed by robustly aggregating local descriptors. In this approach every descriptor is defined by its relative position with respect to a set of reference points (cluster centers) in a d-dimensional space. More precisely, the n cluster centers, or the codebook $\{\mu_1, ..., \mu_n\}$ is computed and each descriptor x_t is assigned to its K-nearest clusters $NN_\gamma^K(x_t)$, where $\{\gamma = 1...K\}$ denotes the rank of a particular nearest cluster. We introduce the following notation: for descriptor x_t, $NN_\gamma^K(x_t)$ returns the cluster index that is rank γ from x_t. In the following paragraph we will discuss several strategies of assigning descriptors to visual words and present our novel rank-based multi assignment.

1) Single Assignment (SA): In SA, each local descriptor x_t is assigned to one nearest cluster (K=1) with an assignment weight τ_{tj} = 1 if $NN_1^1(x_t) = j$ and τ_{tj} = 0 otherwise. The drawback of single assignment is that it leads to high quantization error when matching descriptors are assigned to different clusters, due to inherent variability in the extracted descriptors (noise). Also the population of vectors assigned to each cluster is small, which is not desirable for robust statistical processing.

2) Multiple Assignment (MA): The aforementioned quantization error can be reduced by assigning descriptors to multiple clusters (typically K=2, 3) with constant assignment weight τ_{tj} = 1 if $NN_\gamma^K(x_t) = j$ and τ_{tj} = 0 otherwise. However this approach doesn't take into account that assignments with lower ranks are more stable (in the sense of assignment repeatability) - the reliability and stability decreases as the assignment rank γ increases.

3) Soft Assignment (SoftA): In this method each descriptor x_t is assigned to cluster j with the soft assignment weight (τ_{tj}) given by the posteriori probability (refer to equation 1). While SoftA has been shown to deliver superior results and is generally considered the state-of-the-art, few studies exist on the local descriptor assignment patterns and behavior. We demonstrate later in this section that SoftA has a significant weakness and can be improved in the context of aggregation schemes. One issue is that SoftA often (60% of all the cases) degrades to single assignment. Another problem is that the assignment weights depend on the distances between a descriptor and cluster centers, meaning that the contributions from various descriptors are unbalanced and noise in descriptor values directly impacts the assignment weights. These observations motivate us to introduce our rank-based multiple assignment approach.

4) Rank-based multiple Assignment (RankA): The RankA scheme aims to address the aforementioned drawbacks of the SA, MA and SoftA methods. Firstly, it reduces the assignment error by effectively quantizing descriptors to multiple cluster centers. Secondly, it increases the probability that many clusters have a sizeable population of local descriptors assigned to them. Finally, the descriptors are assigned to K-nearest clusters with stable rank weights leading to a more balanced and reliable global image representation as compared to the MA and SoftA approaches.

In the proposed RankA, each descriptor x_t is quantized to K-nearest cluster centers and the assignment weights used for aggregation are derived from the ranks. In our experiments, K=3 was found to be optimum for many datasets (retrieval results for Holidays dataset are presented in Figure 2(e)). We define assignment weights based on the empirical probability that two descriptors forming a matching pair (inliers) with specific rank are assigned to the same cluster. This probability depends on the proximity of descriptors to the cluster center in feature space, which can be approximated by the assignment rank γ. We expect rank 1 assignments to be more stable than rank 3.

Our procedure to determine the optimal assignment weights includes two steps: (1) finding a set of matching

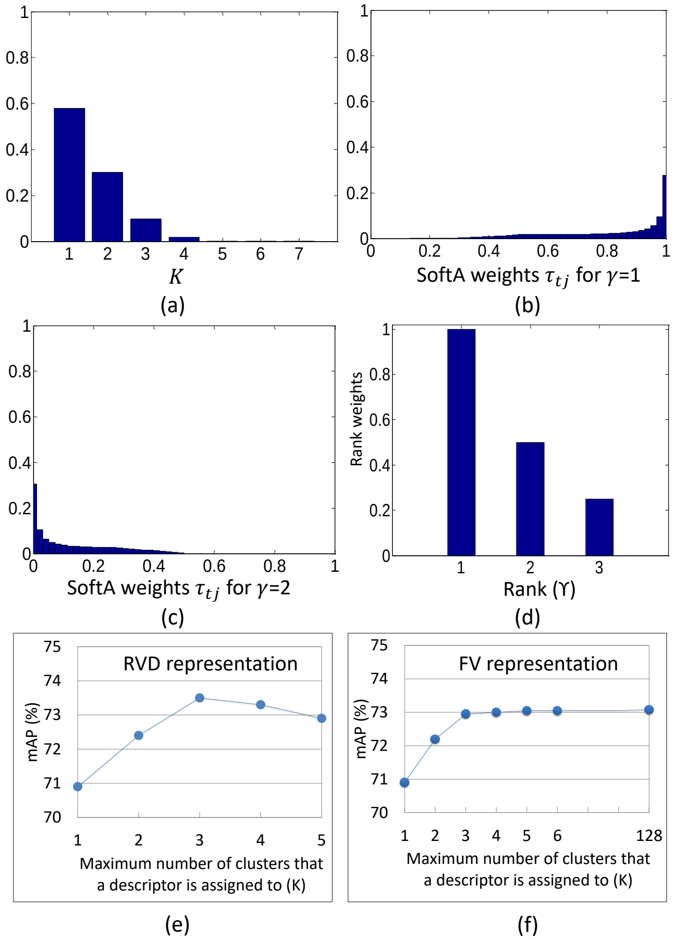


Fig. 2. Fisher Vectors and RVD statistics. The size of codebook is 128, (a) Probability distribution of the number of nearest clusters (K) that a descriptor is assigned to with soft assignment weight greater than 0.1 in FV, (b) Distribution of soft assignment weights corresponding to NN_1^K in FV. About 30% of descriptors are assigned with soft assignment weight of 1, (c) Distribution of soft assignment weights corresponding to NN_2^K in FV, (d) Rank assignment weights used in RVD encoding. In RankA, each descriptor x_t is assigned to three nearest clusters, NN_1^K, NN_2^K and NN_3^K, with assignments weights equal to 1, 0.5 and 0.25 respectively, (e) Performance of RVD as a function of maximum numbers of assigned clusters, (f) Performance of FV as a function of the maximum numbers of assigned clusters.

descriptor pairs and the associated cluster rank-assignment data, and (2) estimating probabilities of the aligned cluster assignment, for each rank γ:

- For a training set of matching image pairs (MPEG dataset [10]), local SIFT descriptors are extracted and a set of putative matches is computed based on Lowe's [1] ratio test. For each image pair a RANSAC algorithm is applied on the putative SIFT matches to estimate an affine transformation and the set of inlier point pairs (Y) consistent with that transform.
- Given inlier point pairs $y_a, y_b \in Y$ and the associated rank-assignment data, we calculate the probabilities (Ω_γ) such that $NN_\gamma^K(y_a) = NN_\gamma^K(y_b)$, for $\gamma = 1..3$. The assignment weights for each γ are calculated as: Ω_γ/Ω_1.

We computed experimentally that in NN_1^K, NN_2^K and NN_3^K, the probability that an inlier point pair is assigned to the same cluster center is approximately $0.58, 0.28$ and 0.14. Therefore assignment weights used in RVD aggregation are: τ_{tj} = 1 if $NN_1^K(x_t) = j$, τ_{tj} = 0.5 if $NN_2^K(x_t) = j$, τ_{tj} = 0.25 if $NN_3^K(x_t) = j$ and τ_{tj} = 0 otherwise.

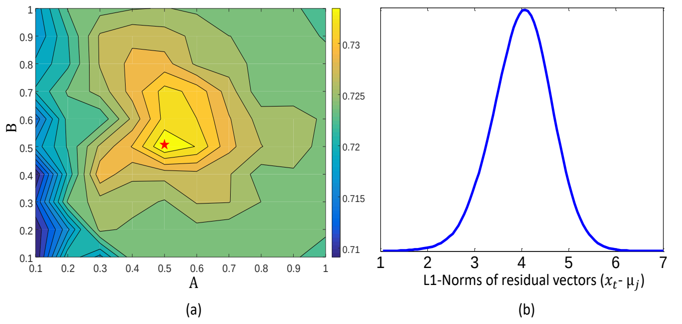


Fig. 3. (a) Grid search for RankA weights (b) Distribution of L1-Norms of residual vectors in RVD scheme.

To confirm that this approach indeed leads to optimal performance, we studied the retrieval accuracy on the Holidays dataset as a function of RankA weights. Since K is set to 3 and the weight corresponding to the nearest neighbor ($\tau_{tj}^{NN_1}$) is set to one, the weights for NN_2 and NN_3 are respectively computed as $\tau_{tj}^{NN_2} = A \times \tau_{tj}^{NN_1}$ and $\tau_{tj}^{NN_3} = B \times \tau_{tj}^{NN_2}$. The values of A and B are varied from 0.1 to 1. The grid search over the aforementioned 2D parameters space (Figure 3(a)) confirms that the weights derived from the assignment statistics are indeed optimal for the RVD representation. The red star indicates the selected RankA parameters based on these statistics. The space exhibits similar behavior for other datasets.

Analysis of SoftA and RankA

In order to better understand the difference between soft assignment (as employed in FV) and our rank-based assignment, we computed FVs for images in the Holidays dataset and analyzed cluster assignment statistics and behavior. The following observations are made; Figure 2(a) shows a discrete probability distribution of the number of nearest clusters (K) that a descriptor is assigned to with soft assignment weight greater than 0.1. It can be observed that in 60% of all the cases the weight assignment in FV is such that only the first nearest cluster has a weight exceeding 0.1. This effectively means that SoftA frequently degrades to single assignment. In RankA, a descriptor is always assigned to three nearest centroids. Figure 2(b) and Figure 2(c) show the distribution of soft assignment weights corresponding to NN_1^K and NN_2^K respectively and it can be seen that in NN_1^K about 30% of descriptors are assigned with a soft assignment weight of 1. In RankA, each descriptor x_t is assigned to three nearest clusters, NN_1^K, NN_2^K and NN_3^K, with assignments weights equal to 1, 0.5 and 0.25 respectively, as shown in Figure 2(d). In Fisher Vector encoding, many assignment weights τ_{tj} are likely to be very small or negligible. We evaluate the performance of FV on the Holidays dataset by setting to zero all but the K-largest assignments for each input descriptor x_t. Figure 2(f) shows that there is no significant change in performance for $K > 3$.

This means that there are no performance benefits arising from using more than three nearest neighbors.

Direction preserving mapping function

Each local descriptor x_t is assigned to its K nearest clusters with corresponding ranks and the residual vectors $x_t - \mu_j$ are calculated. Figure 3(b) shows the distribution of L1-norms of residual vectors $x_t - \mu_j$, where it can be seen that the contribution of individual descriptors to cluster level representation varies significantly. We note that aggregating non-normalized residual vectors leads to suboptimal performance as the cluster level representations can be strongly influenced by outliers with higher magnitudes of residual errors.

To alleviate this problem, we propose that RVD aggregation encodes each local descriptor using only direction, discarding magnitude. More precisely, for each local descriptor and the associated clusters μ_j with ranks γ, the residual vectors $x_t - \mu_j$ are L1-normalized before aggregation. Our choice of L1-norm has been motivated by research showing that in high dimensional spaces the L1-norm exhibits more stable behavior, and is therefore preferable to L2. For example, Aggarwal et al. [28] show that in high dimensional spaces, the concepts of proximity, distance or nearest neighbor may not even be qualitatively meaningful. They examine the behavior of commonly used Lp-norms and show that the problem of meaningfulness in high dimensionality is sensitive to the value of p, with the Manhattan distance being the preferred metric for high dimensional data mining applications. Experimental results on Holidays (Figure 5(a)) and Oxford datasets with varying p coefficient confirm that L1-norm indeed delivers optimal performance.

This direction-preserving mechanism limits the impact of outliers that happen to be located far from a cluster center. In effect, the influence of any single descriptor on the aggregated representative value is now limited and similar in impact to all other descriptors.

RVD formation

Each normalized residual vector belonging to cluster j is weighted based on rank assignment weights τ_{tj} to yield vector $r_{tj} \in \mathbb{R}^d$.

$$r_{tj} = \tau_{tj} \frac{x_t - \mu_j}{||x_t - \mu_j||_1} \tag{9}$$

where $||.||_1$ denote L1-norm.

The cluster level representation $\zeta_j \in \mathbb{R}^d$ is computed by aggregating vectors r_{tj} across all ranks γ.

$$\zeta_j = \sum_{\gamma=1}^{K} \sum_{x_t : NN_{\gamma}^{K}(x_t)=j} \tau_{tj} \frac{x_t - \mu_j}{||x_t - \mu_j||_1} \tag{10}$$

Each ζ_j is L2-normalized (intra-normalization [19]) in order to equalize contributions from all aggregated vectors ζ_j to the final RVD representation R. The dimensionality of vector R is $D = d \times n$.

$$R = \left[\frac{\zeta_1}{||\zeta_1||_2}; \frac{\zeta_2}{||\zeta_2||_2}; ...; \frac{\zeta_n}{||\zeta_n||_2} \right] \tag{11}$$

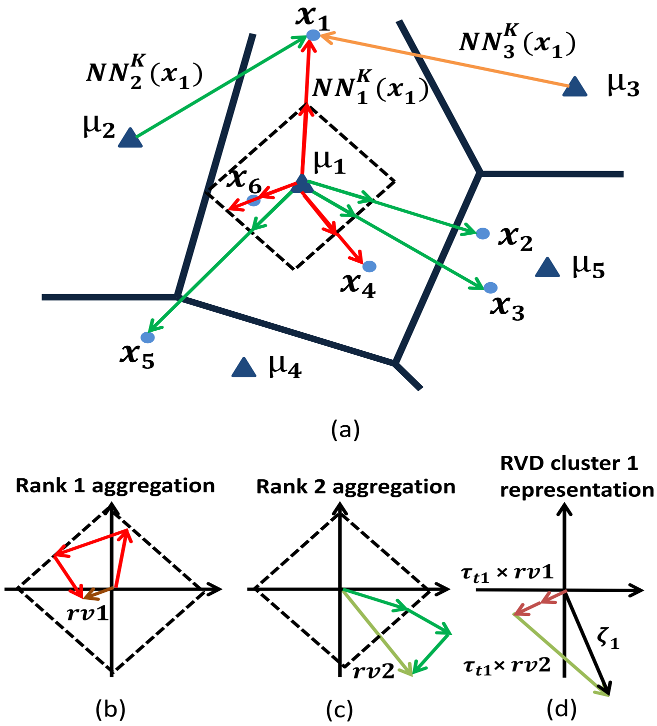


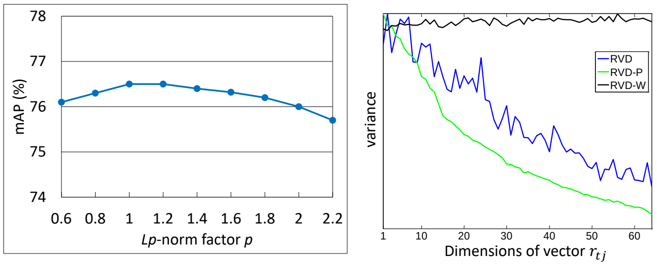
Fig. 4. RVD aggregation approach: (a) rank-based cluster assignment and L1-normalization of residual vectors, (b) Aggregation of residual vectors belonging to Rank-1 of cluster 1, (c) Aggregation of residual vectors belonging to Rank-2 of cluster 1, and (d) RVD cluster-level representation ζ_1.

The final vector R is L2-normalized to make the representation invariant to the number of local descriptors extracted from each image. Furthermore, this also balances the energy of the aggregated vectors between clusters.

An example of the RVD aggregation is shown in Figure 4. The solid polygons indicate Voronoi cells. There are six local descriptors x_1,..,x_6 and five cluster centers μ_1,..,μ_5. The descriptor x_1 is assigned to its three nearest clusters centers (μ_1, μ_2, μ_3), and the corresponding residual vectors, $(x_1 - \mu_1)$, $(x_1 - \mu_2)$, $(x_1 - \mu_3)$, are respectively shown as red, green and orange arrows. The descriptors $\{x_1, x_4, x_6\}$ are assigned to their first nearest cluster (μ_1) and the residual vectors ($x_t - \mu_1$) are L1-normalized (shown by scaling the red colored residual vectors to dashed unit square) in order to discard the magnitude information. The normalized residual vectors are aggregated into rv_1 as shown in Figure 4(b). Similarly, descriptors $\{x_2, x_3, x_5\}$ are quantized to their second nearest cluster μ_1 and the L1-normalized residual vectors (shown by green arrows) are aggregated into $rv2$ (Figure 4(c)). Finally, $rv1$ and $rv2$ are combined with rank assignment weights τ_{tj} into RVD cluster level representation ζ_1 (shown in Figure 4(d)). For simplicity, the figure only shows two ranks.

3.1 Improved RVD

Here we propose two extensions which increase the discriminatory power of the RVD signatures. The first one decorrelates residual vectors r_{tj} by applying cluster-level PCA before aggregation - (RVD-P). The second method aims to balance the variances in different dimensions of individual



(a) (b)

Fig. 5. (a) Holidays performance as a function of Lp normalization applied to residual errors, (b) Energy distribution in each dimension of residual vectors r_{tj} before aggregation into RVD, RVD-P and RVD-W respectively.

residual vectors r_{tj} after the PCA transformation; it is called RVD with local Whitening (RVD-W).

RVD Local PCA (RVD-P)

We improved the performance of RVD signature by transforming the weighted residual vectors $r_{tj} = \tau_{tj}\frac{x_t-\mu_j}{||x_t-\mu_j||_1}$ inside each cluster using a local PCA basis P_j before aggregation into RVD-P. In the following, we describe the process of computing RVD-P representation.

Off-line stage: Given a set of N weighted residual vectors $r_{1j}, r_{2j}, ..., r_{Nj}$ in $\mathbb{R}^d$ extracted from training images, we compute the mean vector $\eta_j = \mathbb{E}[r_{tj}]$ and the covariance matrix Σj for each cluster j.

$$\eta_j = \frac{1}{N_j}\sum_I \sum_{\gamma=1}^{K} \sum_{x_t:NN_\gamma^K(x_t)=j} r_{tj} \tag{12}$$

$$\Sigma_j = \frac{1}{N_j}\sum_I \sum_{\gamma=1}^{K} \sum_{x_t:NN_\gamma^K(x_t)=j} (r_{tj}-\eta_j)(r_{tj}-\eta_j)^\top \tag{13}$$

For each cluster j, we compute a PCA matrix P_j whose columns consists of the orthonormal eigenvectors of Σ_j corresponding to the d largest eigenvalues $\lambda_1 \geq \lambda_2 ... \geq \lambda_d$.

On-line Stage: Given an image I, the vectors r_{tj} are extracted for each cluster j as in the core RVD method. The mean subtracted r_{tj} vectors are projected using P_j before aggregation into cluster level representation ζ_j.

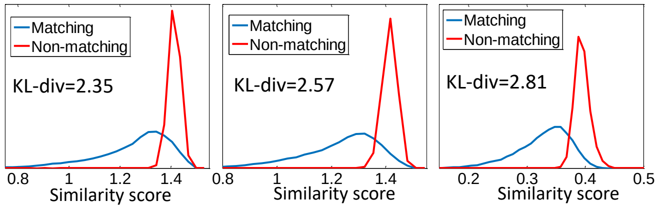
$$\zeta_j = \sum_{\gamma=1}^{K} \sum_{x_t:NN_\gamma^K(x_t)=j} {P_j}^\top(r_{tj}-\eta_j) \tag{14}$$

The final RVD-P representation R^p is formed by concatenating the L2-normalized ζ_j vectors for all clusters.

RVD Local Whitening (RVD-W)

Figure 5 shows the energy distribution in each dimension of residual vectors r_{tj} before aggregation into RVD (blue line), and it can be observed that the variances of different dimensions are not balanced, which negatively affects the discriminability of the final global representation. We solve the aforementioned problem by introducing whitening of the residual vectors r_{tj} before aggregation into cluster level representation.

More precisely, we compute the cluster level whitening matrix P_j^w as P_j^w=$P_j{\Lambda_j}^{-\frac{1}{2}}$, where $\Lambda_j = diag(\lambda_1, \lambda_2 ..., \lambda_d)$.



(a) (b) (c)

Fig. 6. Histogram of Euclidean similarity between matching and non-matching descriptors, for three post-PCA normalization methods (a) Whitening (b) P-L2 (β=0.5) (c) L1-P (β=0.7).

Given an image I, the vectors r_{tj} are computed for each cluster j. The mean subtracted r_{tj} vectors are then projected using P_j and subsequently whitened before aggregation into ζ_j.

$$\zeta_j = \sum_{\gamma=1}^{K} \sum_{x_t:NN_\gamma^K(x_t)=j} {P_j^w}^\top(r_{tj}-\eta_j) \tag{15}$$

The L2-normalized ζ_j vectors are stacked to form the final RVD-W representation R^w.

It can be observed from Figure 5(b) that in RVD-P aggregation, the application of local PCA on r_{tj} concentrates the energy in the top few dimensions while in RVD-W, after performing PCA+Whitening the energy remains balanced between dimensions.

3.2 PCA transformation and L1+Power normalization

In order to improve the separability between matching and non-matching representations, we propose a new normalization approach applied after transforming the RVD-W vectors via PCA. Our normalization involves an L1-norm followed by a power-norm creating L1-P normalization. We show that the L1-P is different to the frequently used Whitening [12] and Power+L2 normalization and offers significant gains in terms of retrieval accuracy.

1) **Whitening**: In [12], Jegou et al. applied whitening operation on the VLAD vector in order to increases the contrast between matching and non-matching descriptors. We follow [12] and perform whitening on RVD-W vector to evaluate its impact on retrieval performance. More precisely, the mean-centered R^w vector is first PCA-transformed, and subsequently whitened and re-normalized to form vector R^{wl}.

$$R^{wl} = \frac{diag(\lambda_1^{-1/2}, .., \lambda_{D'}^{-1/2})P^\top(R^w - R_0)}{||diag(\lambda_1^{-1/2}, .., \lambda_{D'}^{-1/2})P^\top(R^w - R_0)||_2} \tag{16}$$

where R_0 is the mean of the signatures of R^w and P is a $D \times D'$ matrix ($D' \leq D$) of eigenvectors associated with the largest eigenvalues of the covariance matrix of signatures of R^w.

2) **Power+L2 normalization (P-L2)**: The whitening of RVD-W vectors is only suitable when generating short signatures because the smallest eigenvalues produce artifacts. Figure 11(c) demonstrates that the retrieval accuracy initially increases up to 512 dimensions but then decreases when the dimensionality of the RVD-W vector exceeds 512. [13] addressed this problem by applying power-normalization on the PCA projected descriptor, followed

by L2-normalization. The power-norm is parametrized by a constant β.

3) **L1+Power normalization (L1-P)**: In our approach, the mean-centered R^w vector is first transformed using matrix P and then the resultant vector is L1-normalized to form $R^{w\rho}$.

$$R^{w\rho} = \frac{P^\top \times (R^w - R_0)}{||P^\top \times (R^w - R_0)||_1} \tag{17}$$

Finally, the vector $R^{w\rho} = (R_1^{w\rho}, .., R_{D'}^{w\rho})$ is processed using power-normalization: $R_i^{wl} = sign(R_i^{w\rho})|R_i^{w\rho}|^\beta$.

We use the class-separability between matching and non-matching descriptors to demonstrate the advantage of our approach, on MPEG dataset (10k matching and 100k non-matching image pairs) [10]. More precisely, the dimensionality of R^w is reduced to 512 and post-PCA normalization is applied. Let us denote $Pr(h|m)$ and $Pr(h|nm)$ as the probability density function (pdf) of observing a Euclidean distance h for a matching and non-matching descriptor pair respectively. The distance between matching/non-matching pdfs is expressed in terms of KL-divergence. It can be observed from Figure 6 that L1-P method provides the best separability (maximum KL-Divergence) between matching and non-matching distributions, compared to Whitening and P-L2 approaches.

3.3 Compact RVD-W code

The descriptor size, expressed as bytes per image, has a major impact on the performance of an image retrieval system; ideally the descriptors for the entire dataset should fit in the RAM memory of the server for fast processing. Aggregating a 128-dimensional local descriptor (e.g. SIFT) using a small codebook of 64 cluster centers results in 8k-dimensional global descriptor. This size is too large for efficient retrieval in very large databases.

We followed [27] to compress RVD-W vectors into small codes for large scale retrieval. More precisely, the dimensions of vector R^{wp} are permuted using the Eigenvalue Allocation method [27]. The transformed vector is divided into g sub-vectors or groups of equal length D'/g. Each sub-vector is quantized using a separate K-means quantizer with n centroids (256) and encoded using $k = log_2(n)$ bits. The storage requirement of the embedded vector is $B = g \times k$ bits. The distance between the query vector and database vectors is computed using Asymmetric Distance Computation (ADC).

3.4 RVD-W based on Convolutional Neural Networks (CNN)

Recent research has shown that image descriptors computed using deep CNNs achieve state-of-the-art performance for image retrieval and classification tasks. Babenko et al. [14] aggregated deep convolutional descriptors to form global image representations: FV, Temb and SPoC. The SPoC signature is obtained by sum-pooling of the deep features. Razavian et al. [29] compute an image representation by the max pooling aggregation of the last convolutional layer.

We propose to encode CNN-based descriptors into the RVD-W representation. More precisely, an RGB image is first warped into a $c \times c$ square and a mean RGB value is subtracted from each pixel. The image is then passed through a pre-trained network comprising of L convolutional layers. The output of a l-th layer L^l is a $c^l \times c^l \times d^l$ feature map, where d^l is the number of filters corresponding to L^l. A set $X^l = \{x_{1,1}^l, x_{1,2}^l, .., x_{c^l,c^l}^l\}$ of d^l-dimensional feature vectors is obtained at each location (a, b), $1 \leq a \leq c^l$ and $1 \leq b \leq c^l$, in the feature map. As in the SIFT-based approaches, a codebook $\{\mu_1^l, ..., \mu_n^l\}$ of n cluster centers is learned using a set of training images. For each centroid, the residual vectors $x_{a,b}^l - \mu_j^l$ are computed, normalized and whitened to form vector ζ_j^l (Equation 15), regarding layer L^l. The RVD-W representation is obtained by concatenating all aggregated vectors ζ_j^l for all n visual words.

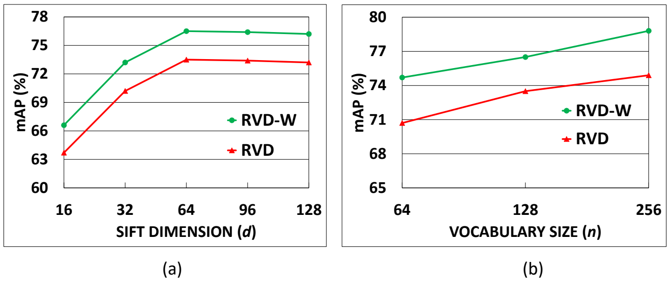


Fig. 7. Impact of parameters on the Holidays performance for RVD-W and RVD representations (a) as a function of SIFT dimensions d and (b) as a function of vocabulary size n (all results in mAP(%)).

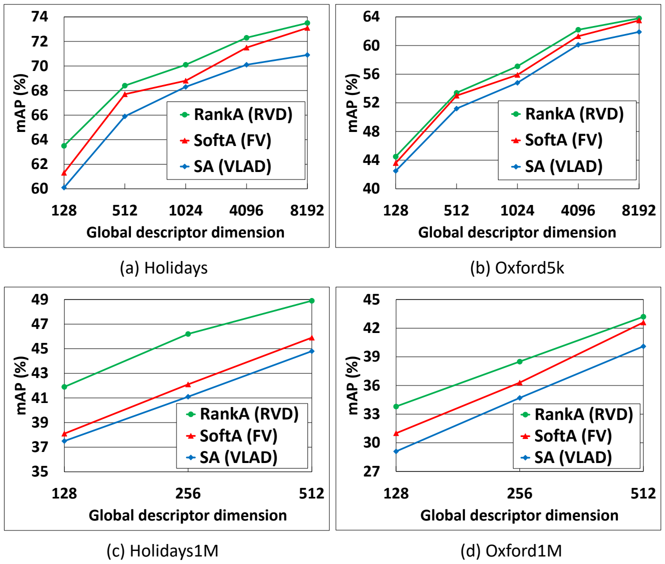
4 EXPERIMENTS

The purpose of this section is to evaluate the RVD-W relative to other state-of-the-art global image representations. We first present the experimental setup which includes the datasets and evaluation protocols. Furthermore, we also define common conditions concerning local descriptor extraction, dimensionality reduction of local descriptors and selection of vocabulary size. We then analyze, based on SIFT features, the impact of the novel components that constitute our method, namely rank-based multiple assignment, the direction preserving mapping function, neighborhood rank weights, application of local PCA and whitening and normalization of the RVD-W vector. Finally, we show that the aggregation framework developed using SIFT features is also effective for CNN-based features. A comparison with the different global representations is presented at the end of this section.

Datasets

The performance of the proposed method is extensively evaluated on three standard image retrieval benchmarks.

The **INRIA Holidays** dataset [30] contains 1491 holiday photos with a subset of 500 used as queries. Retrieval accuracy is measured by mean Average Precision (mAP), as defined in [16]. To evaluate system performance in a more challenging retrieval scenario, the Holidays dataset is augmented with 1 million distractor images obtained from Flickr, forming Holidays1M [30]. We also further extend Holidays1M with additional 9M distractor images [31], to test the robustness of our framework in a very large scale



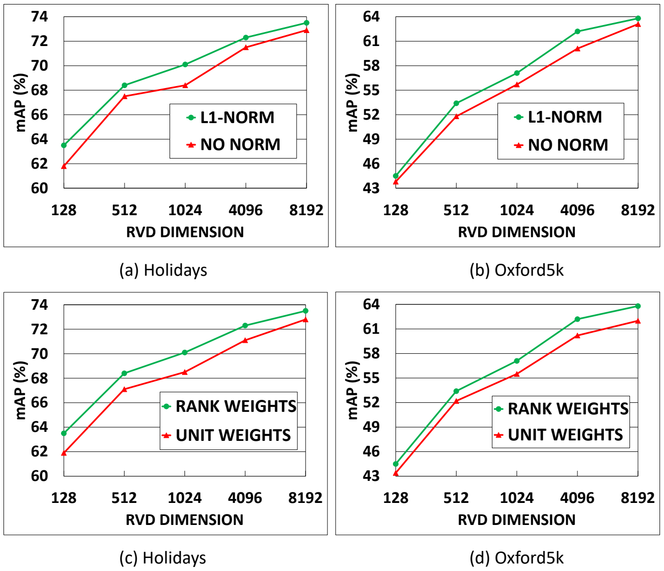
(a) Holidays

(b) Oxford5k

(c) Holidays1M

(d) Oxford1M

Fig. 8. Impact of rank-based aggregation on performance for (a) Holidays, (b) Oxford5k, (c) Holidays1M and (d) Oxford1M (all results in mAP(%))



(a) Holidays

(b) Oxford5k

(c) Holidays

(d) Oxford5k

Fig. 9. Impact of L1-normalization of residual vectors on performance of (a) Holidays dataset, (b) Oxford5k dataset. Impact of rank-based weighting on performance of (c) Holidays dataset, (d) Oxford5k dataset

case. The PCA transformation matrix and visual vocabulary is trained on the Flickr60K dataset [30].

The **University of Kentucky Benchmark** (UKB) [15] dataset contains 10200 images of 2550 objects (4 images of each object). The performance measure is the average number of images returned in the first 4 positions (4 × Recall@4).

The **Oxford5k** dataset [16] contains 5062 images gathered from Flickr by querying for particular Oxford landmarks. From this set of images, 11 distinctive landmarks are selected, with 5 distinct queries per landmark. The performance is evaluated using mAP. Unless stated otherwise, the SIFT descriptors, for the query images, are extracted from inside the ROIs. To test large scale retrieval, this dataset is combined with 100k and 1 million Flickr images [9], forming the Oxford105k [16] and Oxford1M dataset respectively. The Oxford1M dataset is also augmented with 9M distractor images [31] forming Oxford10M dataset. We have used the Paris6k dataset [17] for learning of parameters (PCA and vocabulary).

Local descriptor extraction

In all our experiments, key-points are detected using the Hessian affine detector [32] and local regions are encoded in a 128-dimensional SIFT descriptor [1]. We use the publicly available SIFT descriptors [30] for Holidays and Holidays1M datasets; while for Oxford datasets, the detector and the SIFT descriptors are computed as in [19]. Descriptors are extracted from the UKB and ImageNET datasets (Holidays10M) using software available on-line [8]. The SIFT descriptors are converted to RootSIFT [33] without any additional storage or memory.

The dimensionality of the RootSIFT descriptors is reduced from 128 to d dimensions using PCA matrix. It can be observed from Figure 7(a) that applying PCA and truncating the last 64 dimensions provides the optimum performance for both RVD and RVD-W representation.

Vocabulary size

In this experiment, the impact of vocabulary size on the retrieval performance of RVD and RVD-W was studied. It can be observed from Figure 7(b) that the performance increases as we increase the number of centroids. For $n = 256$ RVD-W results in a mAP=77.2% on the Holidays dataset. However, for higher values of n>256, the size of the global descriptor becomes prohibitive for large scale experiments. In all the following experiments, the size of codebook is fixed at 128 to provide a good trade-off between performance, extraction speed, complexity and memory use.

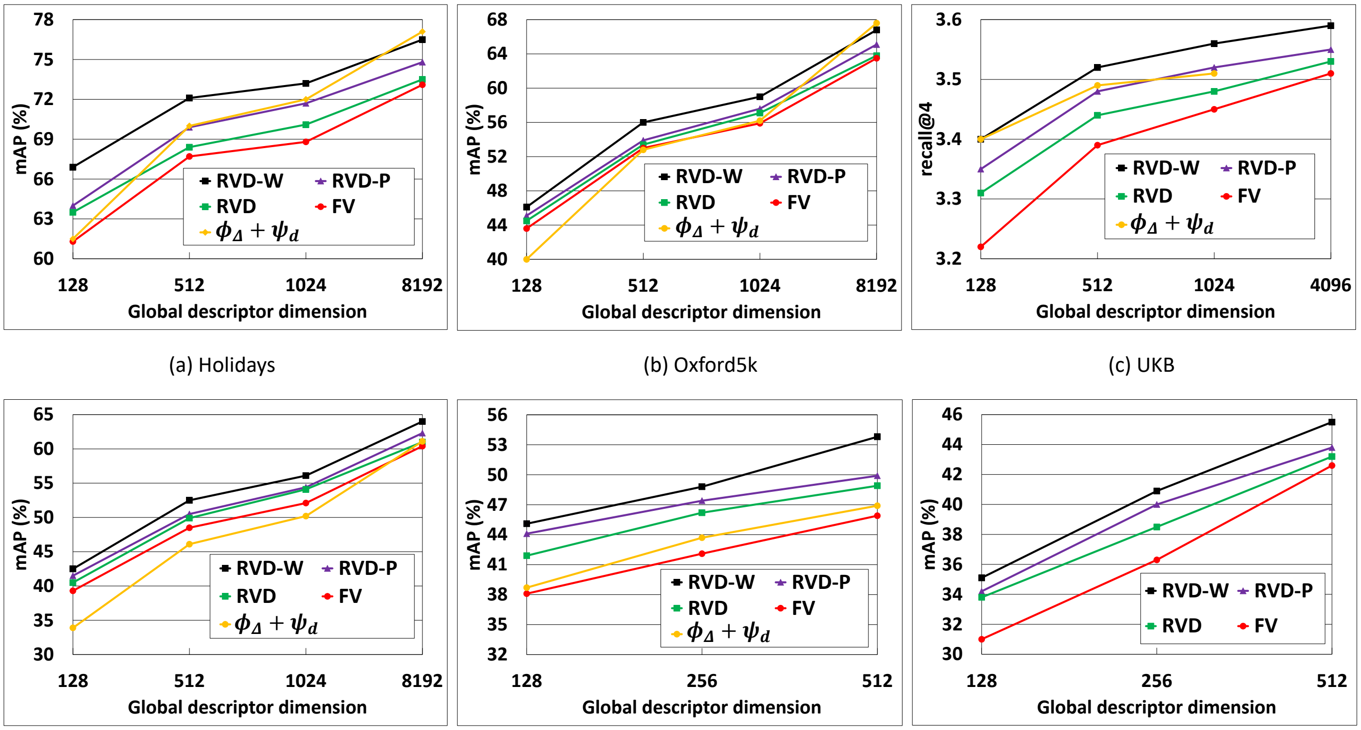
Comparison of descriptor assignment methods

In this section we evaluate the performance of our rank based multiple assignment (RankA) used in RVD, single assignment (SA) employed in VLAD and soft assignment (SoftA) employed in FV, as a function of global descriptor dimensionality D. All global representations are first projected using a $D \times D'$ PCA matrix and then L1-P normalization is applied. The similarity between two descriptors is computed using standard Euclidean distance. It can be seen from Figure 8 that RankA performs better than SA and SoftA approaches on all datasets. Compared to SoftA, the retrieval accuracy obtained using RankA is significantly higher on large scale datasets, resulting in an average gain of 3.6% and 1.9% on Holidays1M and Oxford1M respectively.

Impact of direction-preserving mapping function and neighborhood rank weighting

Figure 9(a) and Figure 9(b) shows the benefit of applying L1-normalization on residual vectors before aggregation. The use of L1-normalization brings an average gain of 1.1% and 1.3% in mAP on the Holidays and Oxford5k datasets respectively.

We performed experiments to show the advantage of neighborhood rank weighting in RVD aggregation process.



(a) Holidays (b) Oxford5k (c) UKB

(d) Oxford105k (e) Holidays1M (f) Oxford1M

Fig. 10. RVD-W comparison with RVD-P, RVD, FV and $\phi_\Delta + \psi_d$, (a) Holidays, (b) Oxford5k, (c) UKB, (d) Oxford105k, (e) Holidays1M and (f) Oxford1M (all results in mAP(%) except for recall@4 for UKB);

In the RVD representation, the weights are: $1, 0.5$ and 0.25 for the assignments with rank one, two and three respectively. It can be observed in Figure 9(c) and Figure 9(d) that weighted rank level combination gives an average improvement of 1.3% and 1.5% on the retrieval accuracy of Holidays and Oxford5k datasets compared to rank level combination with equal weights (as employed in MA approach).

Comparison of RVD/RVD-P/RVD-W/FV/TEmb

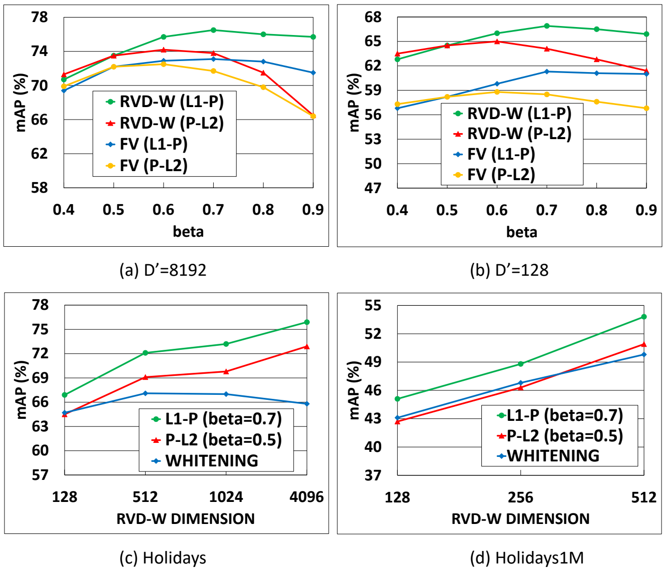
In this section we compare the best representation RVD-W with RVD, RVD-P, FV and TEmb ($\phi_\Delta + \psi_d$). It can be clearly seen from Figure 10 that RVD-W on average outperforms all global descriptors. Compared to FV, RVD-W offers an average gain of +4.5% and +3% in mAP on the Holidays and Oxford5k datasets. The average difference in retrieval performance is even more significant on large scale datasets of Holidays1M (+7%) and Oxford1M (+3.9%) compared to FV. We also compared RVD-W with the recent $\phi_\Delta + \psi_d$ representation. It can be observed that $\phi_\Delta + \psi_d$ obtains marginally better mAP than RVD-W on Holidays and Oxford datasets using a 8192 dimensional descriptor. However $\phi_\Delta + \psi_d$ descriptor suffers significantly from dimensionality reduction and also the computation of $\phi_\Delta + \psi_d$ is typically three orders of magnitude slower than RVD-W. The retrieval performance of RVD-W is significantly better than $\phi_\Delta + \psi_d$ (+5.4% and +6% on Holidays and Oxford5k) after the global descriptors are dimensionally reduced to D'=128. On large scale dataset of Holidays1M, RVD-W offers a significant gain of +6% in mAP over $\phi_\Delta + \psi_d$.

PCA transformation and L1-P normalization

In this section we study how the power-normalization exponent β of P-L2 and L1-P normalizations, effects the retrieval performance of RVD-W and FV. From Figure 11(a) and Figure 11(b), it can be observed that L1-P normalization ($\beta = 0.7$), provides close to optimum performance for both large dimensional (D′=8192) and small dimensional (D′=128) RVD-W descriptor. It is interesting to note that similar behavior is also shown by the FV representation. We performed experiments to compare the performance of the three post PCA normalization methods: (i) Whitening, (ii) P-L2 normalization ($\beta = 0.5$), and (iii) L1-P normalization ($\beta = 0.7$). It can be clearly seen from Figure 11(c) and Figure 11(d) that normalizing the PCA-projected vector using L1-P normalization provides better retrieval accuracy on both the Holidays and the Holidays1M datasets.

Optimized Product Quantization

The purpose of this section is to evaluate the RVD-W representation when used with the joint dimensionality reduction and OPQ method of Section 3.3. The dimensionality of the global descriptor is reduced from 8192 to 128 using matrix P'. The truncated descriptor is L1-P normalized and finally the PQ algorithm is applied on the normalized vector. In all the experiments, we used g=16 sub-vectors and 8 bits to encode each sub-vector resulting in a small code of 16 bytes. The distance between the query descriptor and the database descriptor is computed using ADC. We repeat each experiment 10 times and report the mean performance.



(a) D'=8192

(b) D'=128

(c) Holidays

(d) Holidays1M

Fig. 11. (a) Impact of power-norm exponent β on Holidays dataset using $D' = 8192$ descriptor. Note that P-L2 (β=0.5) is equal to L1-P (β=0.5), (b) Impact of β on Holidays dataset using $D' = 128$ descriptor. Comparison of post-PCA normalization methods on (c) Holidays and (d) Holidays1M

Table 1 shows the performance of compact RVD, RVD-P, RVD-W and Fisher Vector. It can be seen that RVD-W consistently shows better performance on all datasets, achieving 5% higher mAP on the Holidays1M and the Oxford1M compared to FV.

Large scale experiments

Figure 12(a) and Figure 12(b) display the performance of our method on the large scale datasets of Holidays10M and Oxford10M. The mAP performance is presented as a function of dataset size. We show the results for four cases:

- the RVD-W vector reduced to D'=128 by PCA;
- the Fisher Vector reduced to D'=128 by PCA;
- the RVD-W vector compressed to 16 bytes using the 16×8 PQ scheme;
- the Fisher Vector compressed to 16 bytes;

The retrieval performance demonstrate that the RVD-W representation consistently and significantly outperforms FV for both Oxford10M and Holiday10M datasets, typically by a margin of 6% in mAP. Interestingly, it can also be observed that the performance gap increases as the dataset size grows, particularly for the more difficult Oxford dataset, which indicates that RVD-W is more robust in large-scale retrieval. On the ultra-large-scale dataset of Holidays10M, RVD-W (D'=128) obtains a mAP of 40.5% which significantly outperforms any results published to date. To the best of our knowledge, this is the first time that the retrieval experiments have been performed on the Oxford dataset enlarged to 10M. In order to evaluate the performance of our global descriptor in a retrieval system where a short list of images retrieved by RVD-W is re-ranked using local descriptor matching with geometric verification, we evaluated Recall@L i.e. the number of relevant images retrieved in the top L returns. The results are shown in Figure 12(c) and Figure 12(d) where it can be seen that the RVD-W is better than FV in returning correct matches.

TABLE 1
RVD-W, RVD-P, RVD and FV performance using 16 bytes codes

Method	Holidays	Oxford5k	Hol1M	Oxf1M
FV	56.3	38.1	31.0	24.7
RVD	58.1	39.0	33.4	26.8
RVD-P	59.2	39.7	36.3	27.5
RVD-W	**61.4**	**41.2**	**37.3**	**29.0**

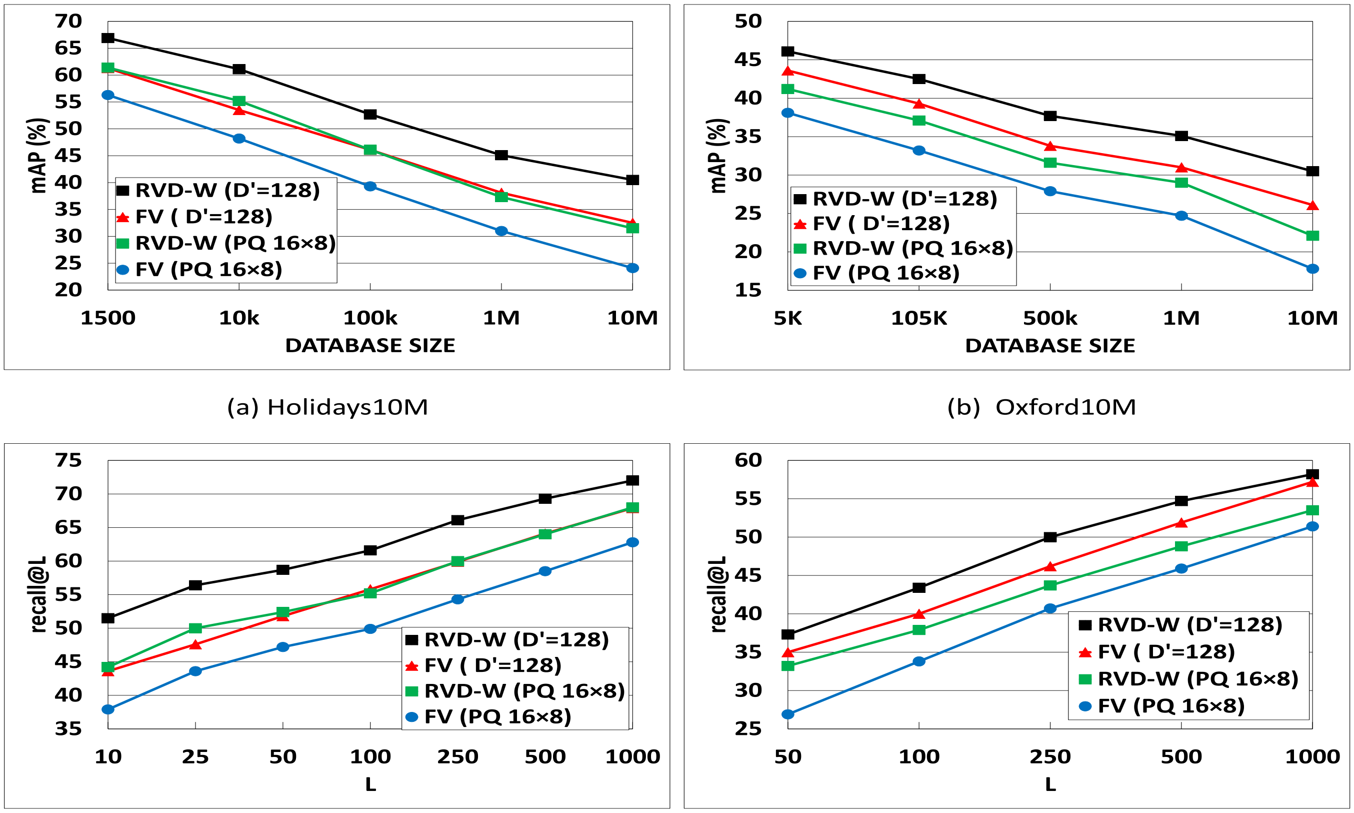
To illustrate the retrieval performance, we compress RVD-W and FV vectors of the Oxford1M dataset, using OPQ, to obtain small codes of 16 bytes. The distance between a query vector and database vectors is computed using ADC, and for every query Recall@100 is calculated. We observe that, out of a total of 55 queries, RVD-W obtains better recall on 20 queries and FV has better recall on 7 queries: RVD-W outperforms FV by 3:1. For an intuitive understanding, Fig. 13 shows three queries where the difference in recall between RVD-W and FV is most significant and one query where the difference in recall between FV and RVD-W is the biggest (maintaining the 3:1 ratio established before). We show the query and the top 4 ranked results obtained by the RVD-W and FV methods using these queries, where correct matches are indicated by a green frame.

5 Comparison with the state of the art

In this section we compare the performance of the proposed method to the latest state-of-the-art algorithms.

Medium footprint image representations (16k-1k dimensions)

Table 2 summarizes the results for medium footprint signatures. In practical applications, the use of medium footprint representations is prohibitive due to search time and memory requirements; however the results are helpful in understanding the capabilities of each representation, and also serve as an upper bound on the expected performance of compact descriptors derived from them. It can be seen that the proposed RVD-W representation outperforms most of the prior-art methods; in particular it improves dramatically (gain of +10% mAP) over the most advanced version of VLAD [20] (referred here as VLAD$_{\text{LCS+RN}}$) and also over FV (gain of +16%), on both Holidays and Oxford databases. Compared to the latest method based on triangulation embedding with sum aggregation ($\phi_\Delta + \psi_s$) [13], RVD-W ($D = 8192$) provides a significant improvement of +3.5%, +2% and +8.5% in mAP on the Oxford, Holidays and Oxford105k datasets. The $\phi_\Delta + \psi_d$ representation performs marginally better than RVD-W on Holidays and Oxford datasets using a 8192 dimensional descriptor. However the $\phi_\Delta + \psi_d$ descriptor suffers significantly from dimensionality reduction as shown by the sharp decrease in performance when the $\phi_\Delta + \psi_d$ descriptor is truncated from 8192 to 1024 dimensions (8k→1k), compared to RVD-W. Also the computation of $\phi_\Delta + \psi_d$ is typically two orders of magnitude slower than RVD-W. On the large scale dataset of Oxford105k, RVD-W offers a gain of +2.9% compared to $\phi_\Delta + \psi_d$. By increasing the number of cluster to 256 the



(a) Holidays10M

(b) Oxford10M

(c) Holidays1M

(d) Oxford1M

Fig. 12. mAP as a function of the database size (a) Holidays10M and (b) Oxford10M. Quality of short-list: recall@L (c) Holidays1M and (d) Oxford1M.

RVD-W (16k) outperforms the $\phi_\Delta + \psi_d$ signature on all datasets. The FAemb (16k) has better retrieval accuracy on Oxford5k compared to RVD-W. However, its extraction time is significantly higher than RVD-W.

Compact image representations

We now focus on a comparison of compact representations which are practicable in large-scale retrieval, as presented in Table 3. The dimensionality of the RVD-W descriptor is reduced from 8192 to 128 via PCA. The results show that our method outperforms all presented methods by a large margin. The gain over the Fisher Vector is $+\mathbf{16}\%$ and $+\mathbf{10}\%$ respectively for Oxford and Holidays datasets. RVD-W provides an improvement of $\mathbf{13.9}\%$ and $\mathbf{16}\%$ on the Oxford5k and Oxford105k datasets over VLAD$_{\text{LCS+RN}}$. Keeping the original descriptor dimensionality of 8192, RVD-W offers gains of $+\mathbf{6}\%$ and $+\mathbf{5}\%$ in mAP on the Oxford and Holiday datasets compared to the $\phi_\Delta + \psi_d$. It should be noted that no results are published for 8192 dimensional $\phi_\Delta + \psi_d$ on Holidays1M dataset because of extremely high encoding times. Compared to $\phi_\Delta + \psi_d$ (D=1920), our method provides an improvement of $\mathbf{6.4}\%$ on Holidays1M. On the ultra large dataset of Holidays10M, RVD-W significantly outperforms the best published results (VLAD+SURF).

Table 4 shows the performance of our method using compact codes obtained by product quantization. Compared to VLAD$_{\text{LCS+RN}}$, the gain remains very significant on Oxford5k ($+\mathbf{14}\%$), Oxford105k ($+\mathbf{16}\%$) and Holidays1M ($+\mathbf{5}\%$). The RVD-W provides a gain of 9.4% on largest Holidays10M dataset over VLAD+SURF.

Compact image representations based on CNN features

This section compares the performance of CNN-based representations suitable for large-scale retrieval. We extract deep convolutional descriptors using the state-of-the-art CNN, OxfordNet [18]. Each image is resized to the size 586×586 before passing through the network. The output of the last layer is a $37 \times 37 \times 512$ feature map, forming a set of 1369 512-dimensional descriptors. We compare RVD-W to the state-of-the art methods successfully used with CNN features: Max-pooling [29], SpoC [14] and FV. We implement SPoC signature (without center prior) following [14]. For RVD-W and FV representations, a codebook of 8 cluster centers are learned via K-means clustering. All methods use final PCA to reduce the global descriptor dimensionality so that it is suitable for large scale retrieval. Training is performed consistently for all methods using the Paris dataset for Oxford, and the Flickr dataset for Holidays. The performance is evaluated using Oxford5k (full query), Holidays, Holidays+1Million and Oxford+1Million datasets. Hol-r denotes a modified Holidays dataset, as used by Babenko et al. [14] (some images are manually rotated). Hol1M is formed by augmenting the original (un-rotated) Holidays images with 1 million distractors. The retrieval performance of the CNN-based representations is presented in Table 5. It can be seen that 256-dim RVD-W improves

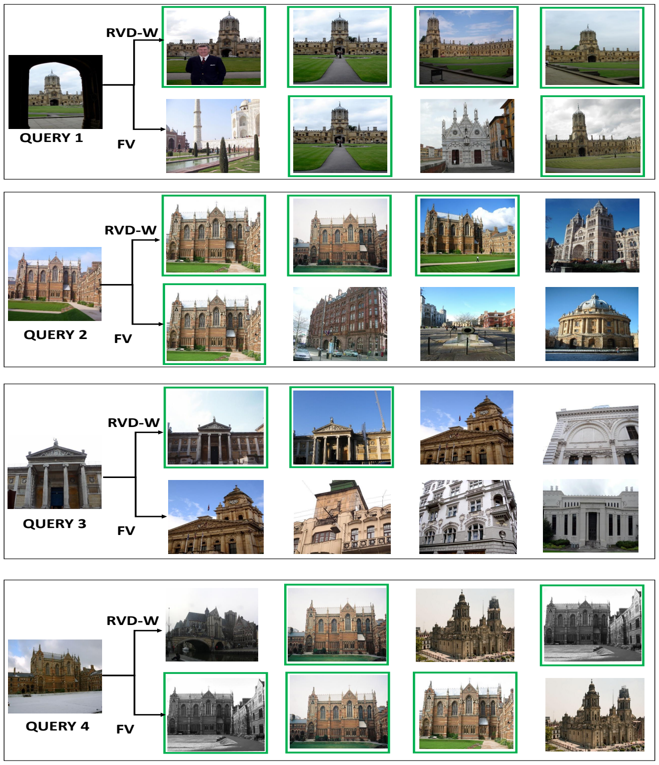


Fig. 13. Example retrieval results for the RVD-W and FV descriptors on the Oxford1M dataset. For each Query image (left) the corresponding ranked lists are shown for the RVD-W (top center-right) and FV (bottom center-right); images correctly retrieved are marked with green border. Both descriptors are quantized using OPQ to 16 bytes.

TABLE 2
Comparison with the state of the art using medium footprint signatures.

Method	Size	Oxf5k	Oxf105k	Hol	UKB
VLAD Intra [19]	32k	55.8	-	65.3	-
HiVLAD [23]	32k	63.8	-	72.1	3.56
VLAD+CSURF [21]	12k	-	-	71.7	3.52
CPVLAT [25]	9k	-	-	70.0	-
VLAD* [20]	8k	50.0	44.5	62.2	-
VLAD$_{\text{LCS+RN}}$ [20]	8k	51.7	45.6	65.8	-
HVLAD [22]	8k	47.2	-	69.1	-
HiVLAD [23]	8k	57.6	-	66.6	3.48
$\phi_\Delta + \psi_d$ [13]	16k	66.5	-	76.8	-
$\phi_\Delta + \psi_d$ [13]	8k	67.6	61.1	77.1	-
$\phi_\Delta + \psi_s$ [13]	8k	63.3	55.5	74.5	-
FAemb [26]	16k	**70.9**	-	78.7	-
FAemb [26]	8k	66.7	-	76.2	-
RVD-W	16k	68.9	**66.0**	**78.8**	**3.60**
RVD-W	8k	66.8	64.0	76.5	3.59
VLAD [8]	4k	37.8	-	55.6	3.28
FV [8]	4k	41.8	-	60.5	3.35
VLAD+SURF [21]	4k	32.8	-	64.9	3.20
$\phi_\Delta + \psi_d$ [13]	8k→1k	56.2	50.2	72.0	-
RVD-W	8k→1k	**59.0**	**56.1**	**73.2**	**3.56**

TABLE 3
Comparison with the state of the art using 96/128 dimensional vectors

Method	Size	Oxf 5k	Oxf 105k	Hol	Hol 1M	Hol 10M
VLAD [8]	128	28.7		55.7	-	-
FV [8]	96	-	-	56.0	31.8	28.0
FV [8]	128	30.1	-	56.5	-	-
VLAD* [20]	128	32.5	26.6	-	33.5	-
VLAD$_{\text{LCS+RN}}$ [20]	128	32.2	26.2	-	39.2	-
CPVLAT [25]	256	-	-	60.6	38.0	-
VLAD+SURF [21]	96	-	-	65.5	42.5	34.0
HiVLAD [23]	128	-	-	64.0	43.0	-
$\phi_\Delta + \psi_d$ [13]	8k→128	40.0	33.9	61.5	▽[1]	▽
$\phi_\Delta + \psi_d$ [13]	2k→128	43.3	35.3	61.7	38.7	-
RVD-W	128	**46.1**	**42.5**	**66.9**	**45.1**	**40.5**

TABLE 4
Comparison with the state of the art with compact codes via PQ

Method	Size	Oxf 5k	Oxf 105k	Hol	Hol 1M	Hol 10M
VLAD [8]	40 B	-	-	49.5	-	-
FV [8]	16 B	-	-	50.6	28.7	21.0
VLAD* [20]	16 B	28.9	22.2	-	29.9	-
VLAD$_{\text{LCS+RN}}$ [20]	16 B	27.0	21.0	-	32.3	-
VLAD+SURF [21]	10 B	-	-	58.0	30.2	22.1
RVD-W	16 B	**41.2**	**37.1**	**61.4**	**37.3**	**31.5**

TABLE 5
Comparison with the state of the art with CNN-based compact codes

Method	Size	Oxf 5k	Hol	Hol-r	Hol 1M	Oxf 1M
MOP-CNN [34]	512	-	78.4	-	-	-
Max-pooling [29]	256	53.3	74.2	-	-	-
SPoC [14]	256	58.9	78.5	80.2	62.2	41.1
FV	256	52.2	76.4	78.1	58.1	35.5
RVD-W	256	**60.0**	**79.7**	**81.3**	**63.5**	**44.8**
MOP-CNN [34]	2048	-	80.2	-	-	-
Max-pooling [29]	2048	58.0	70.7	-	-	-
FV	2048	64.1	81.9	-	-	-
RVD-W	2048	**67.5**	**84.5**	-	-	-

over FV, delivering a gain of +7.8% on Oxford and 3.3% on Holidays. Compared to Max-pooling, RVD-W provides an improvement of +6.7% and 5.5% on Oxford and Holidays datasets. On large scale datasets Hol1M and Oxf1M, RVD-W offers a gain of +1.3% and +3.7% compared to the best performing state-of-the-art SPoC signature. Using cropped queries for Oxford5k (features extracted from ROI), RVD-W 256-Dim achieves 55.5% mAP compared to 53.1% of SPoC.

The 2048-dimensional RVD-W outperforms all CNN-based approaches. It should be noted that the performance of SPoC deteriorates when a 512 dimensional signature is used (79.6% on Holidays and 55% on Oxford).

Razavian et al. achieved 89.6% and 84.3% mAP on Holidays and Oxford dataset using multi-resolution search and Jittering algorithm. The extraction time and matching complexity of their method is prohibitively high for large scale image retrieval. However for the sake of comparison, we performed experiments with RVD-W combined with the Multi-resolution search (3 × 3) and Jittering (3 × 3). The 16k dimensional RVD-W achieves 91.5% mAP on the Holidays dataset. Furthermore, a 256-dim RVD-W signature outperforms their small footprint representation on all datasets.

6 Conclusion

This paper presents a novel method for extraction of a robust and highly discriminative global descriptor called

[1] The symbol ▽ indicates that the experiments could not be performed because the encoding time is prohibitively large.

RVD-W. The key ideas include a novel robust aggregation approach with rank-based multi-assignment, direction-based accumulation, and mid-stage de-correlation and whitening of the residual vectors. The proposed aggregation is also combined and shown to be effective with CNN-based features, outperforming global descriptors based on the sum-pooling approach. A detailed evaluation on defacto standard benchmarks demonstrates that in large-scale retrieval our scheme outperforms state-of-the art methods.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, pp. 91–110, 2004.

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

[3] S. A. J. Winder, G. Hua, and M. Brown, "Picking the best daisy." in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, pp. 145–175, 2001.

[5] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of GIST descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.

[6] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*, 2006.

[7] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors." in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.

[8] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1704–1716, Sep 2012.

[9] M. Bober, S. Husain, S. Paschalakis, and K. Wnukowicz, "Improvements to TM6.0 with a robust visual descriptor proposal from University of Surrey and Visual Atoms," in *MPEG Standardisation contribution : ISO/IEC JTC1/SC29/WG11 CODING OF MOVING PICTURES AND AUDIO, M30311*, jul 2013.

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

[11] M. Bober and S. Husain, "Compact and robust signature for large scale visual search, retrieval and classification," 2015, WO Patent App. PCT/GB2014/052,058.

[12] H. Jégou and O. Chum, "Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening," in *European Conference on Computer Vision*, Oct 2012.

[13] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] A. Babenko and V. S. Lempitsky, "Aggregating deep convolutional features for image retrieval," *CoRR*, 2015.

[15] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2006, pp. 2161–2168.

[16] 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.

[17] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

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

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

[21] E. S. Xioufis, S. Papadopoulos, Y. Kompatsiaris, G. Tsoumakas, and I. P. Vlahavas, "A comprehensive study over VLAD and product quantization in large-scale image retrieval," *IEEE Transactions on Multimedia*, pp. 1713–1728, 2014.

[22] C. Eggert, S. Romberg, and R. Lienhart, "Improving VLAD: hierarchical coding and a refined local coordinate system," in *IEEE International Conference on Image Processing*, 2014, pp. 3018–3022.

[23] Z. Liu, H. Li, W. Zhou, T. Rui, and Q. Tian, "Uniforming residual vector distribution for distinctive image representation," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2015.

[24] D. Picard and P.-H. Gosselin, "Improving Image Similarity With Vectors of Locally Aggregated Tensors," in *IEEE International Conference on Image Processing*, Sep 2011.

[25] R. Negrel, D. Picard, and P.-H. Gosselin, "Web scale image retrieval using compact tensor aggregation of visual descriptors," *IEEE Transactions on Multimedia*, pp. 24–33, Mar 2013.

[26] T.-T. Do, Q. D. Tran, and N.-M. Cheung, "Faemb: A function approximation-based embedding method for image retrieval," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[27] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 744–755, 2014.

[28] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proceedings of the 8th International Conference on Database Theory*, 2001, pp. 420–434.

[29] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Visual instance retrieval with deep convolutional networks," *CoRR*, 2014.

[30] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, feb 2010.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[32] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, 2004.

[33] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[34] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European Conference on Computer Vision*, 2014.

Syed Sameed Husain is a Research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom. Dr. Sameed received the MSc and PhD degrees from University of Surrey, in 2011 and 2016, respectively. His research interest is visual search. Syed Sameed is an inventor of a patent and has made several contributions in MPEG CDVS standard.

Miroslaw Bober is a professor of Video Processing at the University of Surrey, U.K. In 2011 he co-founded Visual Atoms Ltd, a start-up specializing in Visual Search Technologies. Between 1997 and 2011 he headed Mitsubishi Electric Corporate R&D Center Europe (MERCE-UK). Prof. Bober received the MSc and PhD degrees from University of Surrey, in 1991 and 1995, respectively. His research interests include various aspects of computer vision and machine intelligence, with recent focus on image/video database retrieval and data mining. He has been actively involved in the development of MPEG standards for over 20 years, chairing the MPEG-7, CDVS and CVDA groups. Dr. Bober is an inventor of over 70 patents and several of his inventions are deployed in consumer and professional products. His publication record includes over 80 refereed publications, including three books and book chapters.