ROBUST AND SCALABLE AGGREGATION OF LOCAL FEATURES FOR ULTRA LARGE-SCALE RETRIEVAL

Syed Husain and Miroslaw Bober

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK.
e-mail: {s.husain, m.bober}@surrey.ac.uk

ABSTRACT

This paper is concerned with design of a compact, binary and scalable image representation that is easy to compute, fast to match and delivers beyond state-of-the-art performance in visual recognition of objects, buildings and scenes. A novel descriptor is proposed which combines rank-based multi-assignment with robust aggregation framework and cluster/bit selection mechanisms for size scalability. Extensive performance evaluation is presented, including experiments within the state-of-the art pipeline developed by the MPEG group standardising Compact Descriptors for Visual Search (CVDS).

Index Terms— Visual Search, Compact descriptors, Local descriptor aggregation.

1. INTRODUCTION

Modern visual search systems, in particular mobile ones, require algorithms with high recognition performance, supporting scalable and compact bitstream representations. Additional requirements include (1) low computational complexity to improve execution speeds and battery life, and (2) low system memory to reduce silicon costs. ISO/MPEG is currently standardising Compact Descriptors for Visual Search (CDVS) [1] and it has formulated requirements based on inputs from industry, focusing on a scalable descriptor with the size ranging from 512Bytes to 16kBytes per image and implementable with system memory below 128kB. This paper is concerned with the design of a compact, binary and scalable image representation that is easy to compute, fast to match and which delivers beyond state-of-the-art performance in visual recognition of objects, buildings and scenes. In this paper we present a novel scalable descriptor, called Robust Visual Descriptor (RVD). RVD combines a novel aggregation scheme, conceptually based on a combination of robust statistics with rank-based assignment [2, 3]. We additionally introduce bitrate scalability by employing cluster selection and bit selection mechanisms which support interoperable binary image representations that can be easily adapted to the requirements of the use scenario, e.g. communication channel bandwidth or storage limitation. Before we introduce our method we briefly review key prior art.

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

Related work. Much work has been done on how to best aggregate local descriptors (BoW, FV, VLAD) and their binary representations, with continuous performance improvements.

Bag-of-Words (BoW) [4] represents each image as a histogram of visual words. The tf-idf weighting is typically applied and inverted list is used for fast retrieval. Although the BoW method provides reasonable retrieval performance, it produces a sparse representation resulting in large descritptors. It offers limited size scalability and the search time and memory requirements become prohibitive for large databases (above 1 million). To address these issues, several alternative global representations were proposed recently, where local descriptors are first assigned into a relatively small visual vocabulary (typically 64-512 clusters) followed by an aggregation/encoding stage. Perronnin et al [5] applied Fisher kernel to aggregate local image descriptors into compact vector representation (Fisher Vectors - FV). This method assumes a parametric generative model for local descriptors, typically Gaussian Mixture Model (GMM), and model parameters are trained off-line. FV is constructed by aggregating gradients of descriptors log-likelihoods with respect to the model parameters. In [6], Jegou et al proposed a vector of locally aggregated descriptors (VLAD), which builds an image representation by aggregating residual errors for the grouped descriptors based on a locality criterion in the feature space.

Several notable improvements have been made recently to the original VLAD representation. In [7], Arandjelovic introduced Intra-normalisation, tiling images with multiple VLAD descriptors (MultiVLAD) and the Vocabulary adaptation technique, improving retrieval accuracy. Delhumeau [8] achieved significant improvement by introducing two complementary techniques to VLAD: Residual Normalisation and Local Coordinate System via PCA; local coordinate system was also suggested in [9] using LDA.

Since we are interested in compact, binary representations, the following techniques are relevant. Perronnin et al [5] compressed FV (CFV) by using sign binarisation and Hashing techniques. Chen et al [9] introduced Residual

Enhanced Visual Vectors (REVV), where global descriptor dimensionality is first reduced using linear discriminative analysis (LDA) and then binarisation is performed. The drawback of CFV and REVV is that they are not scalable. Jegou et al [6] produced compact image representation by applying PCA and product quantisation (PQ). However, PQ and PCA require large codebooks and PCA matrix consuming considerable memory. Recently Lin et al [10], introduced rate-adaptive compact fisher codes (RCFC) for visual search based on scalable FV representation, demonstrating good performance.

This paper introduces a compact, binary and scalable image representation based on a new robust aggregation approach combined with cluster and bit selection mechanisms. The main premise behind RVD is that the aggregation mechanism should be designed to tolerate large number of outliers (i.e. local descriptors present only in one image due to e.g. occlusion, missing background, etc.). The proposed image representation is easy to extract, fast to match and delivers beyond state-of-the-art performance in visual recognition.

This paper is organized as follows. The pipeline design for the Robust Visual Descriptor is presented is Section 2, followed by implementation details in Section 3. Section 4 describes the evaluation protocols and databases and shows detailed experimental results which demonstrate that our approach significantly outperforms the state-of-the-art methods. Finally, conclusions are presented in Section 5.

2. RVD SCALABLE PIPELINE

The scalable RVD pipeline is presented in Figure 1. Firstly, SIFT descriptors are extracted from an image and their dimensionality is reduced via PCA. The compressed descriptors are rank-assigned to multiple visual words and a robust representation of residual errors in each cluster is derived forming RVD global descriptor. RVD is compressed using Word selection and Bit selection technique.

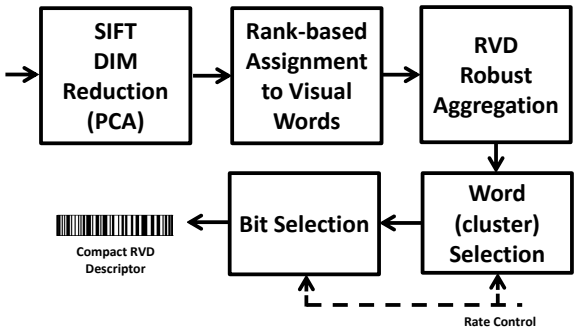

Fig. 1. RVD pipeline

Rank-based assignment. K-means Clustering is performed to learn a codebook $K = \{k_1, ...k_n\}$ of n cluster centres, typically between 64 and 512. The codebook size is selected to provide a good trade-off between the performance, extraction speed, complexity and memory use. For each local descriptor, the distances to all centres are computed and ranked in increasing order. Each descriptor is then rank-assigned to KN closest visual words (typically KN=3) and the rank information is subsequently used for aggregation. By increasing the number of vectors assigned to each cluster and using the rank information for aggregation our scheme delivers improved performance. Note that our assignment weight depends only on the rank information (NR) independently of the actual distance from the respective cluster centres as in the case of soft assignment. The rank based approach also differs from multiple assignment with equal weights because rank information guides the aggregation process. In the VLAD aggregation context, multiple assignment has been recently used [11] but only on the query side and for large numbers of clusters (65k). In our robust framework we note that rank-based multiple assignment brings significant benefits especially in binary domain (see Section 4).

Robust aggregation. We have recently developed a novel approach to aggregation [2], based on the concepts from Robust Statistics. The central idea is that the aggregated representations should match even if only a small proportion of the component local descriptors, corresponding to the objects or regions present in both images, are matching. In other words, aggregated representation should be robust to outliers (i.e. local vectors that do not match). Since it is not possible to determine *apriori* which are the matching ones, it follows that the aggregation scheme should be designed so that all residual vectors have the same influence on the aggregated representation. In the conventional VLAD representation, local vectors located further from the class centres have larger influence, which is something we would like to avoid. The robustness objective is achieved by applying a L1 normalisation to the residual vectors before aggregation, which limits the influence of outliers on the representation. In the example on Figure 2, local vectors $\{x_1, x_4, x_6\}$ belong to rank-1 neighbourhood NR1 of C_1 and their corresponding residual vectors are L1 normalised before aggregation into $rv1$, as shown in (b). Similarly, local vectors $\{x_2, x_3, x_5\}$ belonging to NR2 of C_1 are separately aggregated into $rv2$, as shown in (c). Subsequently, $rv1$ and $rv2$ are combined with weights 2 and 1, as shown in (d). Further details of the aggregation mechanism and study of its behaviour are be provided in [3].

Scalability via cluster selection and bit selection mechanisms. The cluster occupancy and rank are used to estimate reliability of each cluster level representations, which is used to select a subset of clusters with high reliability from the image-level RVD descriptor and additionally also used for rate control of the produced RVD representation. A particular cluster is rejected if the number of local descriptors assigned to that cluster is less than cluster selection threshold C_{th}. The threshold value C_{th} is selected based on typical (me-

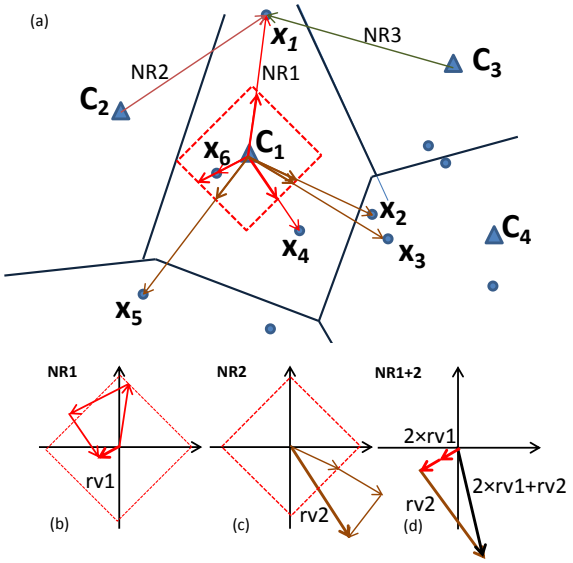

Fig. 2. RVD aggregation approach: (a) rank-based cluster assignment and L1 normalisation, (b) Rank-1 aggregation (NR1), (c) Rank-2 aggregation, and (d) combination of ranks.

dian) cluster occupancy to achieve the required size of RVD descriptor for each bitrate. The RVD vector is binarised by applying the sign function which assigns the value 1 to any non-negative values, and the value 0 to any negative values, respectively. To further compress RVD a subset of bits from the aforementioned binary representation is selected from each cluster, based on separability criteria. We select those bits which provide best separability between hamming distances for matching and non-matching pairs (trained off-line) of binary component-level RVD descriptors. Let $P(X/M)$ and $P(X/N)$ denote the conditional probability that the XOR between two corresponding bits is 1 for matching and non-matching image pairs respectively. We select bits that maximise the difference between $P(X/M)$ and $P(X/N)$. The RVD descriptor size is 1 kilobyte which can be scaled down to any required bitrate via cluster selection and bit selection mechanisms. In MPEG TM7 four interoperable global descriptor sizes were used: 288B, 341B, 460B and 760B, which after addition of the compressed local SIFT descriptors and their locations created overall descriptor sizes between 512B and 16kB.

Low memory. Table 1 compares the memory requirement of RVD, CFV, REVV, and RCFC. For RVD, the SIFT PCA matrix is 128×48 (1 byte per component), plus a 128-dimensional mean vector (1 byte per dimension). The table containing Cluster Centres is 170 × 48 elements (1 byte per element) and a bit selection table is (2×170×48 bits). Overall memory requirement for RVD is just over 16kB compared to 119kB of REVV and 49kB of RCFC.

Method	SIFT (PCA)	Vocabulary	Descriptor transform	Total
REVV	-	24kB (KM)	95kB (LDA)	119kB
CFV RCFC[10]	17kB	33kB (GMM)	-	49kB
RVD	6kB	8kB (KM)	2kB (BSel)	16kB

Table 1. Memory footprint of REVV, CFV, RCFC and RVD (KM:k-means; GMM: Gaussian Mixture Model; BSel: Bit selection)

3. IMPLEMENTATION DETAILS

For MPEG CDVS experiments, we follow TM [1]: (1) all images are resized (max side <= 640), (2) SIFT features are extracted using VLFeat library and a subset of 300 local descriptors is selected based on confidence factor, (3) SIFT dimension is reduced to 48 via PCA and, (4) the size of visual vocabulary is 170. To compare with the state of the art on Holidays, Oxford and UKB datasets, we follow the experimental scenario [7]: (1) for Oxford 5k the detector and SIFT descriptors are computed as in [12] while for Holidays we use publically available SIFT descriptors as in [6], (2) SIFT descriptors are converted to RootSIFT [13], (3) SIFT dimensionality is reduced to 64-dim using PCA and (4) the number of cluster centres used is 128 resulting in 8192 dimensional RVD vector. For parameter training we also followed [7] : we trained on Paris6k [7] for Oxford5k experiments and on Flickr10k for Holidays experiments. In all experiments, we used KN = 3, and rank weights $4, 2, 1$ as it provides close to optimum results.

Fast binary matching. We employ a very fast matching algorithm based on Hamming distance. Given two binary descriptors u^x, u^y extracted from query image X and database image Y, the similarity score $S_{X,Y}$ is specified as weighted correlation score:

$$S_{X,Y} = \sum_{i=1}^{n} b_i^X b_i^Y w_{Ha}(u_i^X, u_i^Y) + (E_1 \times P_1) + (E_2 \times P_2) \quad (1)$$

We have $b_i^X = 1$ if the i^{th} cluster is used for representation, otherwise $b_i^X = 0$. $Ha(.,.)$ denotes the Hamming distance and w_{Ha} denotes the weights to the Hamming distance. Weights w are learned from matching/non-matching image pairs from an independent dataset. E_2 represents the number of times a particular cluster is present in image X and absent in image Y and P_1 is the penalty associated with it. E_2 represents the number of times a particular cluster is absent in both images X and Y and P_2 is the penalty associated with it. The constant penalty values $P_1 = -0.2$ and $P_2 = 0$ were found experimentally to perform well for all bitrates.

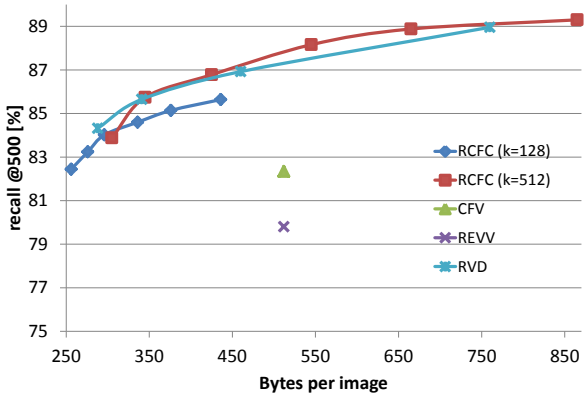

Fig. 3. Recall ©500 at different bitrates, 1M distractors [%].

4. PERFORMANCE EVALUATION

An extensive evaluation has been performed following MPEG CDVS test protocols and additionally using independent datasets: Holidays, UKBench and Oxford. The CDVS evaluation consists of pairwise image matching and retrieval experiments. Five image categories are used [number of query/database images]: (1) Text and graphics including Book/DVD covers/documents/ business cards [1500/1000], (2) Photographs of Paintings [400/100], (3) Video frames [400/100], (4) Landmarks [3499/9599] and (5) Common objects from the UKbench dataset [2550/2550]. Additional datasets include Holidays [1491/500], UKBench and Oxford [5062/55], which are commonly used for benchmarking. The performance is evaluated in terms of Mean Average Precision (mAP) and Recall @ N (N=500) i.e. the number of relevant images retrieved in top N returns. In order to evaluate performance for large scale retrieval, a set of 1million distractors collected from FLICKR is used (FLICKR1M).

4.1. Evaluation on MPEG CDVS

Figure 3 shows the average retrieval performance in terms of Recall @ 500 plotted against different bit rates over MPEG CDVS datasets including 1M destractors. RVD significantly outperforms REVV, CFV and RCFC (k=128). RVD achieved the performance of RCFC (k=512) using only 170 clusters compared to 512 clusters used by RCFC. Also note that RVD memory footprint is only 16kB compared to 49 kB of RCFC.

RVD performance is also evaluated within MPEG CDVS TestModel 7.0 [1]. The TM7.0 framework performs weak geometric verification, based on logarithmic distance ratio (LDR), on the shortlist of images retrieved by RVD. The results show that RVD performance is significantly superior to RCFC and REVV. On average across all datasets, we obtain mAP 83.32% compared to RCFC 81.47% and REVV 77.04%.

Method	Dim	Size	OX 5k	Hol	UKB
BoW	20k	10kB	35.4	43.7	2.87
BoW [6]	200k	12kB	36.4	54.0	2.81
VLAD [6]	8k	32kB	37.8	55.6	3.28
FV [6]	8k	32kB	41.8	60.5	**3.35**
VLAD Intra [8]	8k	32kB	44.8	56.5	-
VLAD* [8]	8k	**32kB**	**50.0**	**62.2**	-
BoW Binary [14]	20k	8.3kB	-	45.8	3.02
FV Binary [5]	8k	1kB	-	58	3.25
FV Binary [5]	4k	0.5kB	-	57.4	3.21
RVD Binary	8k	**1kB**	**50.3**	**61**	**3.34**
RVD Binary	4k	0.5kB	44.3	58.9	3.28

Table 2. Performance comparison with state-of-the-art methods [mAP] and descriptor size [kB]; top 6 methods are non-binary representations, bottom 5 are binary representations.

4.2. Evaluation on Holiday, Oxford and UKB datasets

Table 2 compares binary RVD representation with leading methods on Oxford, Holidays and UKB dataset. Compared to state-of-the-art binary Fisher Vector (bottom part of the table), RVD shows consistent and significant improvement on all datasets, in fact RVD even at 0.5kB outperforms FV at the 1kB operating point. To get a full picture of relative performance, we also show results for non-binary representations of the same dimensionality. These are at least an order of magnitude larger (float vs bit per dimension) and significantly slower in matching (Hamming popcount vs L2 norm). We note that even in this case RVD outperforms all reference methods, except for the case of the Holiday dataset where non-binary VLAD* representation achieves 62.2% vs binary RVD 61% (non-binary RVD achieves 67.8% in a comparable pipeline [3]). We treat the non-binary results for reference only as they do not conform to the low-memory requirements outlined in the introduction. In order to assess the benefits of rank-based assignment we experimented with single-assignment case, which resulted in 3.4% drop in mAP performance on the Holiday dataset. For KN=3 multiple assignment with same weights, the performance drop was 1%.

5. CONCLUSIONS

A novel global image descriptor is proposed which combines rank-based assignment with robust aggregation framework and cluster/bit selection mechanisms for size scalability. Extensive experiments demonstrate excellent recognition performance, outperforming latest state-of-the-art algorithms with binary representations and achieving better or comparable performance to non-binary descriptors (at a fraction of their descriptor size). The extraction process requires low memory and matching is very fast, conforming to MPEG CDVS requirements.

6. REFERENCES

[1] "CFP for compact descriptors for visual search," *ISO/IEC JTC1/SC29/WG11/N12201*, 2011.

[2] M. Bober, S. Husain, S. Paschalakis, and K. Wnukowicz, "Improvements to TM6 with a Robust Visual Descriptor - Proposal from University of Surrey and Visual Atoms," *ISO/IEC JTC1/SC29/WG11 MPEG2013/M30311*, July 2013.

[3] S. Husain and M. Bober, "On Robust Aggregation of local image descriptors (preprint)," , 2014.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003, vol. 2, pp. 1470–1477.

[5] F. Perronnin, Y. Liu, J. Snchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors.," in *CVPR*. 2010, pp. 3384–3391, IEEE.

[6] 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*, vol. 34, no. 9, pp. 1704–1716, Sept. 2012.

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

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

[9] David Chen, Sam Tsai, Vijay Chandrasekhar, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Process.*, vol. 93, no. 8, pp. 2316–2327, Aug. 2013.

[10] J. Lin, L. Duan, T. Huang, and W. Gao, "Rate-adaptive Compact Fisher Codes for Mobile Visual Search," *IEEE Signal Processing Letters (to appear)*, 2014.

[11] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou, "To aggregate or not to aggregate: selective match kernels for image search," in *ICCV - International Conference on Computer Vision*, Sydney, Australia, Sept. 2013.

[12] M. Perďoch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 9–16.

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

[14] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Packing bag-of-features," in *IEEE International Conference on Computer Vision*, Sep 2009, pp. 2357–2364.