Rate-adaptive Compact Fisher Codes for Mobile Visual Search

Jie Lin, Ling-Yu Duan, Yaping Huang, Siwei Luo, Tiejun Huang, and Wen Gao

***Abstract*—Extraction and transmission of compact descriptors are of great importance for next-generation mobile visual search applications. Existing visual descriptor techniques mainly compress visual features into compact codes of fixed bit rate, which is not adaptive to the bandwidth fluctuation in wireless environment. In this letter, we propose a Rate-adaptive Compact Fisher Codes (RCFC) to produce a bit rate scalable image signature. In particular, RCFC supports fast matching of descriptors based on Hamming distance; meanwhile, low memory footprint is offered. Extensive evaluation over benchmark databases shows that RCFC significantly outperforms the state-of-the-art and provides a promising descriptor scalability in terms of bit rates versus desired search performance.**

***Index Terms*—Compact descriptors, compression, Fisher vector, mobile visual search, rate adaptation.**

I. Introduction

CAMERA equipped mobile devices are becoming ubiquitous platforms which facilitate mobile visual search (MVS) applications like Google Goggles. Existing MVS applications may search rich objects, such as CD/book cover, poster, logo, landmark, product, etc. A system usually transmits query images from a mobile client to a remote server, then performs visual query over a reference image database hosted on the server. In wireless environment, the query response latency depends on the network bandwidth. It often takes a few seconds to transmit a JPEG image ($30 \sim 40$ kB) as a query over a slow link [1]. An alternate approach is to extract visual features directly on a mobile client, compress the features and send compact descriptors to the server over the network [2], [3]. This alternative is expected to significantly reduce network latency and improve user experience. On the other hand, compact descriptors allow fast descriptor matching as well as light storage of visual descriptors extracted from reference database images at the server end. In particular, this topic relates to an ongoing MPEG standardization, namely, Compact Descriptors for Visual Search (CDVS) [4].

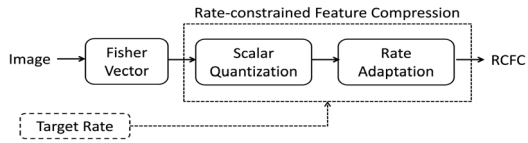

Fig. 1. Framework of rate-adaptive compact Fisher codes.

The development of compact descriptors in MPEG CDVS has to address a key issue of scalability. Generally speaking, existing state-of-the-art compact descriptors [5], [6], [7], [8] follow a typical pipeline: statistics of local invariant features (such as SIFT [9] and SURF [10]) are aggregated to form a fixed-length vector representation, which is subsequently compressed into compact codes for efficient and effective matching, transmission, as well as for significantly reduced memory complexity. These methods encode a high dimensional signature to a short descriptor at hundreds of bytes. However, these methods produce compact image signatures in a fixed bit rate, which is not adaptive to the bandwidth variation in wireless network. Ideally, a small image signature is desirable over a low bandwidth connection (like 2G or 2.5G), while the bit rate of an image signature can be moderately increased to fulfill more discriminative power when more bandwidth is available (like 3G or WiFi). For instance, six operating points $\{0.5\text{ kB}, 1\text{ kB}, 2\text{ kB}, 4\text{ kB}, 8\text{ kB}, 16\text{ kB}\}$ are defined in the ongoing MPEG CDVS standardization [11].

To accommodate the bandwidth variation, a compact image signature is required to support bit rate scalability. In this letter, we formulate the bit rate scalable descriptor coding as a rate-constrained feature compression problem. We develop a Rate-adaptive Compact Fisher Codes (RCFC) [12][1]. Specifically, a rate-adaptive scalar quantization is proposed to compress a fixed-length Fisher vector (FV) representation [5] into binary codes of variable size (see Fig. 1). RCFC supports fast matching between compact descriptors encoded at different bit rates, in which the features of Hamming distance computing and light memory footprint are hardware friendly. To evaluate RCFC, we first study the retrieval performance by using the RCFC signatures compressed at a range of different bit rates. In addition, we employ the evaluation framework of MPEG CDVS, in which the RCFC is combined with geometric verification to improve the retrieval performance. Comprehensive results show that the RCFC not only outperforms the state-of-the-art, but also addresses a balance issue between descriptor bit rates and search accuracy.

Related Work. The Bag-of-Words (BoW) [13], [14] is the most widely adopted image signature for visual search, which

Manuscript received June 24, 2013; revised November 07, 2013; accepted December 16, 2013. Date of publication January 02, 2014; date of current version January 08, 2014. This work was supported by the Chinese Natural Science Foundation under Grants 61272354, 61271311, 61121002, and 61273364. This work was performed at the Institute of Digital Media, Peking University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marco Mattavelli.

J. Lin is with the Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China, and also with the Institute of Digital Media, the School of EE & CS, Peking University (e-mail: jie.dellinger@gmail.com).

L.-Y. Duan, T. Huang, and W. Gao are with the Institute of Digital Media, the School of EE & CS, Peking University, Beijing 100871, China (e-mail: lingyu@pku.edu.cn; tjhuang@pku.edu.cn; wgao@pku.edu.cn).

Y. Huang and S. Luo are with the Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: yphuang@bjtu.edu.cn; swluo@bjtu.edu.cn).

(*Corresponding author: L.-Y. Duan*).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2296532

[1]The proposed RCFC has been adopted as the compact global descriptor in the Committee Draft of the emerging MPEG CDVS standard.

counts the number of local features being assigned to quantized visual words in a visual vocabulary. Recently, the FV representation [5] has extended the BoW by computing higher-order statistics of the distribution of local features. Jegou *et al.* [6], [8] proposed the Vector of Locally Aggregated Descriptors (VLAD) to aggregate the visual word residuals, which can be regarded as a non-probabilistic version of FV. Compared to the BoW, both FV and VLAD have achieved better retrieval performance at a much smaller visual vocabulary.

Compression schemes [5], [6], [7], [8] have been proposed to reduce the bit rate of image signatures. Hashing techniques like locality sensitive hashing (LSH) perform worse at low bit rates [5]. Vector quantization (e.g., product quantizer [6], [8]) and dimensionality reduction (e.g., PCA) based schemes require large codebooks or projection matrices, which may not favor memory-constrained mobile devices. Perronnin *et al.* [5] presented the compressed Fisher vector (CFV) by quantizing each dimension of the FV into a single bit based on a sign function, which outperforms LSH at low bit rates. Chen *et al.* [7] introduced the Residual Enhanced Visual Vector (REVV), where linear discriminant analysis (LDA) is employed to reduce the dimensionality of VLAD, followed by sign binarization to generate compact codes. However, neither CFV nor REVV can address the scalability issues; that is, the code size is fixed and not adaptive to the variable bit budget.

II. Problem Statement

In general, higher dimensional image signatures often bring about more discriminative power at the risk of expensive storage and/or transmission cost. Our goal is to compress raw signatures into small codes, without incurring considerable loss of retrieval accuracy. In addition, the encoder allows the code size of compact signatures to be scalable with respect to different operating points. Finally, to fit the hardware design, the encoding process should incur small memory footprint and less computational complexity. Hence, we formulate the bit rate scalable descriptor coding as a rate-constrained feature compression problem. Formally speaking, given a fixed-length signature $\mathbf{g}^X$ for an image X and a compression function Q (e.g., a quantizer), we aim to generate the compressed descriptors $Q(\mathbf{g}^X)$, with distortion D_X (e.g., mean square error) and bit rate R_X. Accordingly, the objective is to minimize the distortion D_X between the raw signature $\mathbf{g}^X$ and the compressed descriptors $Q(\mathbf{g}^X)$, subject to the constraint that the bit rate R_X approaches a target rate R_{budget}:

$$\min_{Q} D_X(Q) \quad s.t. \quad R_X(Q) \to R_{budget} \tag{1}$$

As the bit rate R_X is varied with respect to the target rate R_{budget}, we have to address a challenging issue of matching descriptors across different code sizes. That is, for a query X_q and a database image X_r, if the code sizes of their compressed descriptors $Q(\mathbf{g}^{X_q})$ and $Q(\mathbf{g}^{X_r})$ are different (i.e., $R_{X_q} \neq R_{X_r}$), $Q(\mathbf{g}^{X_q})$ and $Q(\mathbf{g}^{X_r})$ cannot be compared directly using standard metrics (e.g. Hamming distance). Therefore, the compression function Q is supposed to not only fulfill the bit rate scalability, but also allow the similarity measurement between compressed descriptors of different code sizes. For simplicity, we define the similarity S_{X_q,X_r} between images X_q and X_r as:

$$S_{X_q,X_r} = f(Q(\mathbf{g}^{X_q}), Q(\mathbf{g}^{X_r})), \tag{2}$$

where $f(\cdot,\cdot)$ is a distance function to measure the similarity between compressed descriptors.

III. Rate-adaptive Compact Fisher Codes

To address the objective stated in (1), we employ a FV representation, followed by a rate-adaptive scalar quantization to generate bit rate scalable compact Fisher codes RCFC. Furthermore, the RCFC supports Hamming distance based fast matching between compact descriptors of different code sizes.

A. Brief Review of Fisher Vector

Let $X = \{\mathbf{x}_j\}_{j=1}^t$ be a set of d-dimensional local features extracted from image X. Let u_λ with parameters λ be a probability density function that models the generation process of local features. Jaakkola *et al.* [15] proposed to describe X by the gradient vector of the log-likelihood of the image:

$$G_\lambda^X = \nabla_\lambda \log u_\lambda(X). \tag{3}$$

The gradient describes how to update the parameters λ for better fitting the image X. A natural kernel on this gradient is the Fisher kernel [15]: $K(X_q, X_r) = {G_\lambda^{X_q}}' \mathbf{F}_\lambda^{-1} G_\lambda^{X_r}$, where $\mathbf{F}_\lambda = E_{\mathbf{x}\sim u_\lambda}[G_\lambda^{\mathbf{x}} {G_\lambda^{\mathbf{x}}}']$ is the Fisher information matrix [15] of u_λ. $\mathbf{F}_\lambda$ is positive semi-definite, and can be decomposed as $\mathbf{F}_\lambda^{-1} = {\mathbf{L}_\lambda}' \mathbf{L}_\lambda$. $\mathbf{L}_\lambda$ may be considered as a normalization matrix. Hence, $K(X_q, X_r)$ can be rewritten in the form of dot product between normalized gradient vectors $\mathbf{g}$ with:

$$\mathbf{g} = \mathbf{L}_\lambda G_\lambda^X. \tag{4}$$

We refer to $\mathbf{g}$ as the Fisher Vector (FV) of image X.

Specifically, Perronnin *et al.* [5] chose u_λ to be a Gaussian Mixture Model (GMM) with k centroids: $u_\lambda(\mathbf{x}_j) = \sum_{i=1}^k \omega_i p_i(\mathbf{x}_j)$, $\lambda = \{\omega_i, \mu_i, \sigma_i^2\}_{i=1}^k$, where ω_i, μ_i and σ_i^2 are the weight, mean vector and variance vector of the ith Gaussian (We assume diagonal covariance), respectively. The GMM parameters λ are estimated over a training set of local features using the Expectation-Maximization (EM) algorithm. Assuming that the local features are i.i.d., let $\mathbf{g}_i$ be the d-dimensional gradient vector w.r.t. the mean μ_i of the ith Gaussian, we have:

$$\mathbf{g}_i = L_{\mu_i} \frac{\partial \log u_\lambda(X)}{\partial \mu_i} = \frac{1}{\sqrt{t\omega_i}} \sum_{j=1}^{t} \gamma_j(i) \sigma_i^{-1} (\mathbf{x}_j - \mu_i), \tag{5}$$

where $\gamma_j(i) = \omega_i p_i(\mathbf{x}_j) / \sum_{l=1}^k \omega_l p_l(\mathbf{x}_j)$ denotes the probability of local descriptor $\mathbf{x}_j$ being generated by the ith Gaussian. Finally, the FV $\mathbf{g}$ is formed by concatenating the Fisher sub-vectors $[\mathbf{g}_0, \ldots, \mathbf{g}_k]$ of all Gaussians and is therefore kd-dimensional. Readers are referred to [5] for more details.

B. Scalar Quantization

In this work, we choose an one-bit scalar quantizer q to binarize the high dimensional FV $\mathbf{g}$, so that superior retrieval performance with nearly zero memory footprint can be achieved. The goal is to encode the FV $\mathbf{g} \in \mathbb{R}^{kd}$ using a binary vector $\mathbf{b} \in \{-1, 1\}^{kd}$, each element g of the FV $\mathbf{g}$ is projected to 1 if $q(g) > \theta$; otherwise, -1. θ is a threshold. The quantizer q is supposed to minimize the quantization error[2]:

$$D_X(q) = \|\mathbf{g} - \mathbf{b}\|^2. \tag{6}$$

[2]In the formulation, each entry of FV is quantized to $\{-1, 1\}$. In practice, we use $\{0, 1\}$ for the convenience of Hamming distance computation.

Through expanding (6), we obtain the following form:

$$\|\mathbf{g}-\mathbf{b}\|^2 = \|\mathbf{g}\|^2 + kd - 2\mathbf{g}'\mathbf{b}. \quad (7)$$

Thus, minimizing (7) can be solved by maximizing $\mathbf{g}'\mathbf{b}$. Note that the normalized FV representation is nearly zero-centered, i.e., $\sum_{i=1}^{kd} g_i \approx 0$. To maximize $\mathbf{g}'\mathbf{b}$, we simply set the binarized code $b_i = 1$ whenever $g_i > 0$; otherwise, $b_i = -1$. Accordingly, we derive the quantizer $q(g) = sgn(g)$ in the form as:

$$sgn(g) = \begin{cases} 1 & \text{if } g > 0; \\ -1 & \text{otherwise.} \end{cases} \quad (8)$$

For a vector, $sgn(\cdot)$ denotes the element-wise results by (8).

C. Rate Adaptation

With the scalar quantizer q, kd bits are required to encode the FV $\mathbf{g}$. If $kd > R_{budget}$, we opt to quantize Fisher sub-vectors $\{\mathbf{g}_0, \ldots, \mathbf{g}_k\}$ via per Gaussian basis to meet rate constraints. Our empirical observation has shown that the original FV signal presents a sort of Gaussian basis sparsity. Indeed, if none of local feature $\mathbf{x}$ was assigned to Gaussian i, then all the elements of the corresponding Fisher sub-vector $\mathbf{g}_i$ in (5) are zero. Obviously, the sparse vector $\mathbf{g}_i$ is less informative but contributes to the quantization error in (6). To save bit budget, those sparse Fisher sub-vectors can be discarded.

We propose to encode Fisher sub-vectors in a progressive manner to generate binary codes of the FV $\mathbf{g}$, till the bit budget has been fully occupied. Specifically, we have $s_i = 1$ if the Fisher sub-vector $\mathbf{g}_i$ of the ith Gaussian is selected; otherwise, $s_i = 0$. If $s_i = 1$, the quantization error by $\mathbf{g}_i$ is $\|\mathbf{g}_i - sgn(\mathbf{g}_i)\|^2$. The number of selected Gaussian functions equals to the number of non-zero entries of 0/1 vector $\mathbf{s}$, i.e., $\|\mathbf{s}\|_1$, thereby yielding bit rate $R_X = d\|\mathbf{s}\|_1$. Given the quantizer q, the rate-constrained optimization in (1) can be rewritten by extending (6) as follows:

$$\min_{\mathbf{s}} \sum_{i=1}^{k} s_i \|\mathbf{g}_i - sgn(\mathbf{g}_i)\|^2 \quad s.t. \quad d\|\mathbf{s}\|_1 \rightarrow R_{budget}. \quad (9)$$

Given a fixed dimension d, R_{budget} actually determines $\|\mathbf{s}\|_1$. The optimization in (9) can be done efficiently by applying a sorting algorithm to the set $\{\|\mathbf{g}_i - sgn(\mathbf{g}_i)\|^2, i = 1 \ldots k\}$. In other words, the Fisher sub-vector $\mathbf{g}_i$ with the smallest quantization error is first selected to generate Fisher codes, followed by the $\mathbf{g}_j$ with the second smallest quantization error, and so on. The stop criteria is that the bit rate $d\|\mathbf{s}\|_1$ has reached the target rate R_{budget}. The rate-adaptive signature is referred to as RCFC. Accordingly, the resulting RCFC bitstream has an overhead of k bits to keep track of the role of each Gaussian function [12].

Discussion. The memory footprint of RCFC equals to CFV, e.g., $k(1+2d)$ parameters for the GMM. When we remove the rate constraint in (1), the proposed RCFC degenerates to CFV. Experiments show that RCFC not only reduces the bit rate of CFV, but also significantly improves the performance, especially in the context of CDVS Core Experiments [16].

D. Hamming Distance Matching

The proposed RCFC elegantly supports the similarity matching of compact Fisher codes compressed at different code sizes. Given a query X_q and a database image X_r with bit rates $d\|\mathbf{s}^q\|_1$ and $d\|\mathbf{s}^r\|_1$ respectively, the similarity S_{X_q,X_r} in (2) is specified as a normalized cosine similarity score:

$$S_{X_q,X_r} = \frac{\sum_{i=1}^{k} s_i^q s_i^r (d - 2h(sgn(\mathbf{g}_i^{X_q}), sgn(\mathbf{g}_i^{X_r})))}{d\sqrt{\|\mathbf{s}^q\|_1 \|\mathbf{s}^r\|_1}}, \quad (10)$$

TABLE I
THE BIT RATE OF REVV, CFV AND THE PROPOSED RCFC

Method	Target rate (bytes)					
	R_1	R_2	R_3	R_4	R_5	R_6
REVV [7]	512					
CFV [5]	512					
RCFC (k=128)	256	276	296	336	376	436
RCFC (k=512)	305	345	425	545	665	865

where $h(\cdot,\cdot)$ is the Hamming distance between binarized Fisher sub-vectors. If the code sizes of image X_q and X_r are different ($\mathbf{s}^q \neq \mathbf{s}^r$), S_{X_q,X_r} is computed based on the overlapping Gaussians $\mathbf{s}^q \cap \mathbf{s}^r$ between X_q and X_r.

IV. EXPERIMENTS

Datasets and Evaluation Metrics. To evaluate the performance of the proposed RCFC, we carry out retrieval experiments over public available datasets [11]: (1) *Graphics* dataset depicts CD/DVD/book cover, text document and business card. There are 1,500 queries and 1,000 database images; (2) *Painting* dataset contains 400 queries and 100 database images of paintings (say history, portraits, etc.). (3) *Frame* dataset contains 400 queries and 100 database images of video frames captured from a range of video contents like movies and news. (4) *Landmark* dataset contains 3,499 queries and 9,599 database images from building benchmarks, including the ZuBuD dataset, the Turin buildings, the PKUbench, etc. (5) *UKbench* dataset contains 2,550 objects, each containing 4 images taken from different viewpoints. (6) *Holidays* dataset is a collection of 1,491 holiday photos, there are 500 image groups where the first image of each group is used as a query. To fairly evaluate the performance over a large-scale dataset, we use *FLICKR1M* as the distractor dataset [11], containing 1 million distractor images collected from Flickr.

The retrieval performance is measured by mean Average Precision (mAP) and Recall@N ($N = 500$), i.e., the relevant images in top N returns. For *UKbench*, we report the average number N_s of relevant images in top 4 returns as well, which is the most common measure over this dataset [14].

Implementation details. All the images are resized with reduced resolutions (max side ≤ 640 pixels). SIFT features are extracted by the VLFeat library. The dimensionality of raw SIFT is reduced from 128 to $d \in \{32, 64\}$ using Principal Component Analysis (PCA) [5], [8]. We evaluate the performance of RCFC with the number of Gaussians $k \in \{64, 128, 256, 512\}$. The Oxford building and the Caltech building datasets are employed as the independent dataset in all training stages to learn GMM models and PCA projection matrices.

Bit rate scalable descriptors. Table I lists the bit rates of REVV, CFV and the proposed RCFC for query images. Both REVV and CFV employ the fixed-length codes in 512 bytes. In contrast, the code size of RCFC is varied with respect to the target rate R_{budget}, e.g., ranging from 256 bytes (60 Gaussian functions are selected) to 436 bytes when $k = 128$ and $d = 32$. However, we fix the code size of RCFC for database images, i.e., 105 and 300 Gaussian functions are selected in which $k = 128$ and $k = 512$, respectively.

Fig. 2 shows the retrieval results in terms of Recall@500 vs. different bit rates over different datasets. Firstly, the RCFC significantly outperforms the fixed-size CFV and REVV over all datasets at all code sizes (except CFV for the Painting dataset). For $k = 128$, CFV and REVV yield Recall@500 82.31% and 79.73% on average over all datasets, while the RCFC has achieved better Recall@500 84.2% at lower bit

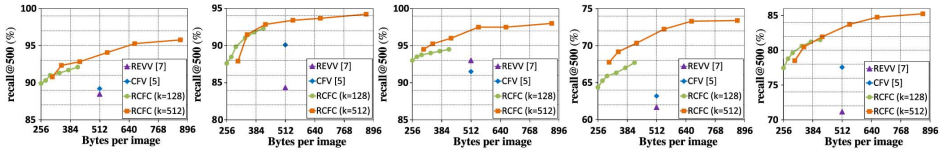

Fig. 2. Recall@500 vs. different bit rates over the various types of datasets, combined with the distractor set FLICKR1M.

TABLE II
COMPARISON OF THE RCFC WITH THE STATE-OF-THE-ART [6], [8] ON UKBENCH (N_s) AND HOLIDAYS (MAP). k DENOTES THE NUMBER OF CENTROIDS

Method	UKbench	Holidays
BoW $k = 20,000$ [8]	2.87	43.7
BoW $k = 200,000$ [8]	2.81	54.0
VLAD $k = 64$, ADC 32×10 [6]	3.10	49.5
FV $k = 64$, ADC 16×8 ($D' = 96$) [8]	3.10	50.6
FV $k = 256$, ADC 256×10 ($D' = 2048$) [8]	3.47	63.4
RCFC $k = 64$	**3.27**	**60.4**
RCFC $k = 256$	**3.46**	**66.7**

TABLE III
MEMORY FOOTPRINT OF REVV, CFV AND THE PROPOSED RCFC

Method	SIFT Transform	Vocabulary	Descriptor Transform	Total
REVV [7]	—	k-means 24kB	LDA 95kB	119kB
CFV [5], RCFC (k=128)	PCA 16.5kB	GMM 32.5kB	—	49kB

rates ($256 \sim 436$ bytes). This gain may be attributed to the fact that RCFC discards less informative Gaussian functions. Secondly, with more bits, RCFC can improve the performance progressively. For example, Recall@500 is increased from 87.91% at 305 bytes to 94.23% at 865 bytes on Painting dataset when $k = 512$. The results have demonstrated the promising scalability of RCFC in seeking a desirable balance between code size and search accuracy.

Comparison with the state-of-the-art. Table II compares the performance of the RCFC with BoW [8], VLAD [6] and FV [8] on two typical benchmark datasets: UKbench and Holidays. The proposed RCFC significantly outperforms BoW. Compared to VLAD [6] and FV [8] with product quantization, the RCFC obtains better search performance with comparable number of centroids. For instance, when $k = 64$, the RCFC achieves a much better mAP 60.4% on Holidays, while VLAD [6] reports 49.5% and FV [8] 50.6%.

Combined with geometric verification. We further evaluate the RCFC performance within the MPEG CDVS evaluation framework [11]: a weak geometric verification (GV) is applied to verify the geometric consistency within a shortlist of 500 database images returned by the RCFC based retrieval. Note that GV works on compressed local features [17], which are hosted by using the remaining bit budget of each operating point (e.g., subtract the RCFC with size $R_1 = 256$ bytes from the lowest operating point 0.5 kB). The results show that RCFC+GV yields much better performance than CFV+GV and REVV+GV. For example, RCFC+GV achieves an average mAP 81.47% versus REVV+GV 77.04% over all datasets. Readers are referred to the MPEG CDVS Input Contribution [12] for more comprehensive results.

Complexity analysis. Table III compares the memory complexity of REVV, CFV and the proposed RCFC. For RCFC and CFV, the SIFT PCA projection matrix size is 128×32, plus a 128-dimensional mean vector, in the format of floating point (4 bytes), yielding the cost of $128 \times 33 \times 4 = 16.5$ kB. The GMM parameters involve a set of $\{\omega_i, \mu_i, \sigma_i^2\}$ for Gaussian i, resulting in $128 \times (1 + 32 + 32) \times 4 = 32.5$ kB. Compared with REVV, both RCFC and CFV incur much less memory.

V. CONCLUSION

We have proposed a discriminative and compact descriptor RCFC by bit rate scalable descriptor coding. RCFC exhibits low computational complexity, and supports fast similarity matching of descriptors encoded at different bit rates. Over extensive benchmarks, RCFC has shown promising search performance. A full-fledged search pipeline involving RCFC based retrieval and fast geometric verification has been validated. Particularly, RCFC has been adopted in the Committee Draft of the ongoing MPEG CDVS standard as a compact global descriptor. More research work on the interoperability of state-of-the-art global descriptors will be included in our future work. In addition, how to incorporate feature selection into more informative aggregation and how to distribute the RCFC indexing structure for scalable visual search [18] are promising research topics.

REFERENCES

[1] B. Girod, V. Chandrasekhar, and D. Chen *et al.*, "Mobile visual search," *IEEE Signal Process. Mag.*, 2011.
[2] R. Ji, L.-Y. Duan, and J. Chen *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, 2011.
[3] V. Chandrasekhar, G. Takacs, and D. Chen *et al.*, "Compressed histogram of gradients: A low-bitrate descriptor," *Int. J. Comput. Vis.*, 2012.
[4] ISO/IEC JTC1/SC29/WG11/N12201, CFP for Compact Descriptors for Visual Search 2011.
[5] F. Perronnin, Y. Liu, and J. Sanchez *et al.*, "Large-Scale image retrieval with compressed fisher vectors," *CVPR*, 2010.
[6] H. Jegou, M. Douze, and C. Schmid *et al.*, "Aggregating local descriptors into a compact image representation," *CVPR*, 2010.
[7] D. Chen and S. Tsai *et al.*, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Process.*, 2012.
[8] H. Jegou, F. Perronnin, and M. Douze *et al.*, "Aggregating local images descriptors into compact codes," *PAMI*, 2012.
[9] D. G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vis.*, 2004.
[10] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *ECCV*, 2006.
[11] ISO/IEC JTC1/SC29/WG11/N12202, Evaluation Framework for Compact Descriptors for Visual Search 2011.
[12] ISO/IEC JTC1/SC29/WG11/M26726, Peking Univ. Response to Core Experiments 1: A Scalable Low-Memory Global Descriptor 2012.
[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," *ICCV*, 2003.
[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *CVPR*, 2006.
[15] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *NIPS*, 1999.
[16] ISO/IEC JTC1/SC29/WG11/N12930, Description of Core Experiments on Compact descriptors for Visual Search 2012.
[17] J. Chen, L.-Y. Duan, and J. Lin *et al.*, "On the interoperability of local descriptors compression," *ICASSP*, 2013.
[18] R. Ji, L.-Y. Duan, and J. Chen *et al.*, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Trans. Multimedia*, 2012.