AN EFFICIENT CODING FRAMEWORK FOR COMPACT DESCRIPTORS EXTRACTED FROM VIDEO SEQUENCE

Zhangshuai Huang[⋆◇], *Ling-Yu Duan*[⋆◇], *Jie Lin*[†], *Shiqi Wang*[○], *Siwei Ma*[⋆◇], *Tiejun Huang*[⋆◇]

[⋆] Institute of Digital Media, School of EE & CS, Peking University, Beijing, 100871, China
[◇] Cooperative Medianet Innovation Center, Shanghai, China
[†] Institute for Infocomm Research, 119613, Singapore
[○] Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada

ABSTRACT

Towards effective and efficient image matching or retrieval tasks, the emerging MPEG standard, named Compact Descriptors for Visual Search (CDVS), has fulfilled compact descriptors for still images, consisting of compressed local and global descriptor. Nevertheless, the frame-level coding of CDVS descriptors from a video sequence does not address the inter-frame redundancy issue, which may consume considerable bandwidth and storage resources. In this work, we propose an efficient coding framework of CDVS descriptors to generate compact descriptors for video sequences. For local descriptors, we propose a multiple reference predictive technique to exploit the temporal correlation of local descriptors and location coordinates over a sequence of frames. To further improve the prediction performance, keypoint tracking is applied to identify temporally repeated keypoints. For global descriptors, a propagation coding way is employed to compress the global descriptors of adjacent frames. The empirical evaluation has shown that the proposed coding approach has yielded a low bit rate of less than 40kbps on average, while maintaining comparable matching and retrieval performance. Compared to the sequence of original frame-level CDVS descriptors, the proposed approach has achieved over 25× bit rate reduction.

Index Terms— Compact descriptor, MPEG CDVS, Interest points tracking, Predictive coding, Propagation coding.

1. INTRODUCTION

Video analysis applications, such as mobile augmented reality, visual sensor network and distributed surveillance, usually transmit visual data from mobile client to remote server for the subsequent matching or retrieval tasks. Instead of sending raw data of images or videos, recent works [1] [2] have proposed to directly extract low bit rate visual descriptors on mobile client, towards low latency delivery in wireless environment. In general, visual descriptors can be broadly categorized into two groups. The first group is local descriptor, such as SIFT [3], SURF [4]. The second group is global descriptor, such as Bag-of-Words [5], Fihser Vector (FV) [6] and Vectors of Locally Aggregated Descriptors (VLAD) [7]. These global descriptors are usually aggregated from the statistics of local descriptors.

Compact representation of local and global descriptors has drawn many research attentions. For compact local descriptor, Chandrasekhar et al. proposed a Compact Histogram of Gradients (CHoG) descriptor with ∼50 bits. Other representative works include BRIEF [8], ORB [9] and BRISK [10]. For compact global descriptor, Chen et al. [11] introduced Residual Enhanced Visual Vector (REVV) by reducing the VLAD dimension with Linear Discriminative Analysis (LDA) followed by sign binarization. Lin et al. [12] proposed Scalable Compressed Fisher Vector (SCFV) to directly binarize FV followed by centroid basis bit selection. In particular, the emerging MPEG standrad [13], Compact Descriptors for Visual Search (CDVS), has standardized both compact local and compact global descriptors. It has shown that CDVS obtains state-of-the-art image matching and retrieval performance at a low bit rate [14].

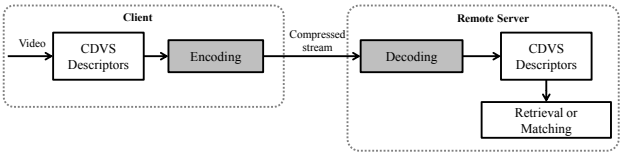

Fig. 1. Overview of our proposed approach. CDVS descriptors extracted from video are encoded at client and transmitted to server for further retrieval or matching task.

Nevertheless, there is few work on compressing descriptors extracted from video sequence, especially for CDVS descriptor. Unlike still image, video is born with the so called temporal redundancy issue. Recent work has proposed to address this issue on either local or global descriptor. For local descriptor, Markar [15] proposed a temporally coherent keypoint detector and inter-frame canonical patches coding techniques. Baroffio [16][17] adopted both intra- and inter-frame coding to compress SIFT- and BRIEF-like [8] descriptors, where a coding mode decision scheme was proposed to improve the coding efficiency. For global descriptor, Chen [18] proposed inter-frame coding of scalable residual-based global signatures REVV by propagating either codewords or residual vectors.

In this paper, we propose an efficient coding framework to compress CDVS descriptors steam. An overview of the our proposed approach is illustrated in Fig.1. In the details of our framework, we investigate a coding pipeline for both local and global descriptors (see Fig.2). We first introduce a tracking process, before the feature selection stage of CDVS framework, to recognize the repeated keypoints for better utilizing the temporal consistency. Then we employ a multiple reference predictive coding technique to reduce the temporal redundancy of local descriptors and location coordinates. At last, an efficient propagation coding technique is designed to compress the global descriptors. Extensive experiments have been conducted over the Stanford MAR Dataset [15]. The results have shown that our approach can achieve a significant bit rate reduction, while with little effect on the image matching and retrieval performance.

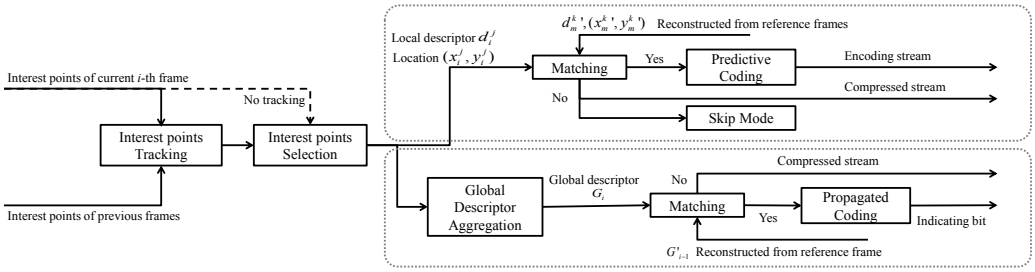

Fig. 2. Block diagram of the proposed coding pipeline. The dotted arrow indicates tracking process is an optional stage.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the MPEG CDVS standard. We present the proposed video descriptors coding framework in Section 3. The evaluation of the proposed scheme is illustrated in Section 4. Finally, Section 5 draws the conclusions and discusses future work.

2. BRIEF REVIEW OF CDVS

MPEG CDVS standard provides a standardized description of still images for efficient and inter-operable visual search applications, such as image matching and retrieval. To ensure compactness and scalability, CDVS has defined six operating points with different descriptor size, say $\{$0.5kB, 1kB, 2kB, 4kB, 8kB, 16kB$\}$.

There are several key processing steps to construct the compact descriptor of an image: (i) A set of local invariant descriptors are extracted from an image by keypoints detection with the 128-Dim SIFT descriptions [3]. (ii) A subset of reliable local descriptors are selected using a pre-trained keypoint selection model [19]. (iii) The selected SIFT descriptors are compressed with transform scalar quantization [20], resulting in a set of binary local descriptors. The size of each binary local descriptor ranges from 40 bits to 256 bits, which is adapted to the defined operating points. (iv) The location coordinates of the selected keypoints are compressed based on their location density map and a histogram count array, followed separately by a simple arithmetic coder and a context based arithmetic coder [21]. The size of each compressed keypoint location ranges from 4.5 bits to 7.5 bits on average. (v) A binary global descriptor, Scalable Compressed Fisher Vector (SCFV), is generated by aggregating the selected uncompressed SIFT descriptors into a single vector, followed by sign binarization. SCFV offers a scalable and compact representation (with an average size from 304 to 1117 bytes). A SCFV descriptor can be denoted as

$$G = \{(s^1, g^1), (s^2, g^2), ..., (s^m, g^m)\}, s \in \{0, 1\} \qquad (1)$$

where s is a binary variable indicating if the i-th Gauss function was selected and generated corresponding binary gradient vector g_i. One can see that the CDVS descriptor is mainly comprised of a set of compressed local descriptors with associated location coordinates and one global descriptor SCFV.

Unfortunately, for video descriptors coding, directly extracting CDVS descriptors frame by frame leads to heavy transmission and storage. For example, assume operating point 4kB is adopted for the CDVS descriptor, the bit rate of descriptors for a video at 30fps is 960 Kbps. Considering the temporal redundancy between video frames, our objective is to design an efficient yet effective video descriptors coding framework that achieves compact descriptors for video representation. In particular, we extend the CDVS descriptor to video domain by injecting temporal correlation into the CDVS framework.

3. CODING SOLUTION

We denote the CDVS descriptor for the i-th frame as $\{D_i, L_i, G_i\}$ where $D_i = \{d_i^j\}$ and $L_i = \{(x_i^j, y_i^j)\}$ refer to a selected subset of compact local descriptor d_i^j and their associated coordinate (x_i^j, y_i^j), G_i refers to the compact global descriptor. In this section, we present the proposed video descriptors coding pipeline (as shown in Figure 2 in a block diagram), consisting of (i) a tracking process to select the repeated (reliable) keypoints; (ii) a predictive coding technique to compress local descriptors and location coordinates between video frames; (iii) a propagation coding scheme to compress the global descriptors between video frames.

3.1. Interest points tracking

Repeated (reliable) keypoints are these interest points appearing in different frames but describing the same visual content. They are important to several visual tasks (e.g. 3D reconstruction, object tracking), and also benefit to the following predictive coding process. In practice, we introduce an interest points tracking process to identify the repeated keypoints among continuous frames by employing a ratio test[3] matching strategy, that is:

$$\frac{\boldsymbol{D}(d_i^j, d_k^m)}{\boldsymbol{D}(d_i^j, d_k^n)} < \ell \qquad (2)$$

where d_i^j is the query descriptor in current frame f_i. d_k^m and d_k^n are nearest and second nearest descriptor in reference frame $\{f_k\}_{k=i-1}^{k=i-N_{win}}$. $\boldsymbol{N_{win}}$ is denoted as the window size of successive frames prior frame f_i. The distances are calculated using Hamming Distance in the compressed domain. When Eq.(2) is satisfied, the d_i^j will be treated as a repeated(reliable) keypoint. After that, the final selected interest points, encapsulated into final CDVS descriptor C_i, are constituted by two groups: one group is repeated keypoints, the other is the interest points selected by the selection strategy. It should be mentioned that tracking process is an optional stage before interest points selection, so that our technique is compatible with the standard MPEG CDVS pipleline.

3.2. Local descriptors predictive coding

In the following, we divide local descriptors coding into two stages: *prediction* and *encoding*.

(i) *Prediction*. For each d_i^j of frame f_i, we employ a matching step as prediction to find a reference local descriptor d_k^m from reference frames. The matched pairs also satisfy the ratio test [3]. Moreover, inspired by the multiple reference strategy in video coding, we try to search the reference local descriptors from multiple

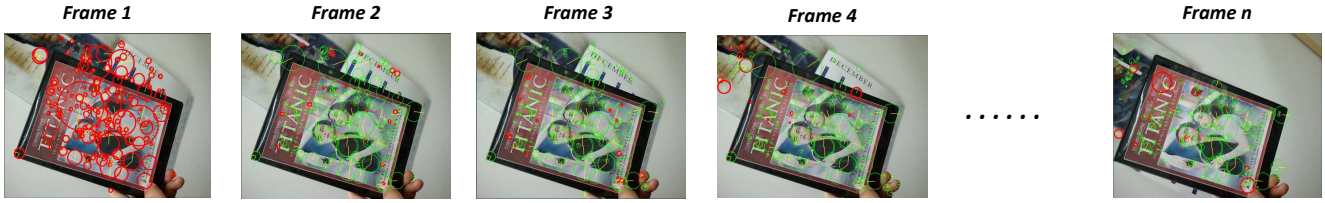

Fig. 3. Best viewed in color version. Keypoints prediction : $I-locals$ are original keypoints in red and $P-locals$ that have reference are keypoints in green.

reference frames. Specifically a predictive function P is defined:

$$D_i = \boldsymbol{P}(D_k), k \in \{i-1, i-2, ..., i-N_{refer}\} \tag{3}$$

Where D_i of i-th frame is predicted from local descriptors D_k of k-th frame, the number of reference frames is denoted as $\boldsymbol{N_{refer}}$.

(ii) *Encoding*. After prediction stage, we categorize local descriptors D_i into two types. A local descriptor d_i^j with a reference local descriptor is called $P-local$, otherwise, called $I-Local$. Fig.3 shows $I-locals$ in red and $P-locals$ in green. For $P-locals$, a d_i^j is predicted from the reconstructed reference $d_k^{m\prime}$, which is decoded from reference frame. So we only need to encode the residual signal between d_i^j and $d_k^{m\prime}$. In practice, we directly set d_i^j=$d_k^{m\prime}$ and discard the residual signal, which is negligible residue energy between two local descriptors from consecutive frames. As a result, we only need to encode a refer index of $d_k^{m\prime}$ for d_i^j, this is proved to achieve a significant bit rate reduction without degrading image matching performance. For $I-Locals$, a huffman coding method adopted in MPEG CDVS standard is performed. To further reduce the bit rate, we introduce a ***Skip Mode*** for every $\boldsymbol{AS}$ successive frames, where we do not encode the $I-Locals$ of the first $(\boldsymbol{AS}-\mathbf{1})$ frames but encode all locals in the $\boldsymbol{AS}$-th frame.

3.3. Location coordinates Coding

Each location coordinate from the set L_i=$\{(x_i^j, y_i^j)\}$ is related to a local descriptor d_i^j. For $\{(x_i^j, y_i^j)\}$ of $I-Locals$, we perform location coding method [21] adopted in MPEG CDVS standard [14], or we discard it when employing ***Skip Mode***. For $\{(x_i^j, y_i^j)\}$ of $P-Locals$, most of them can be predicted from the reference location $\{(x_{i-1}^m, y_{i-1}^m)\}$ in previous frame f_{i-1} using affine transformation. It is calculated as follows:

$$\begin{aligned}\begin{bmatrix} x_i^{j\prime} \\ y_i^{j\prime} \\ 1 \end{bmatrix} &= A \begin{bmatrix} x_{i-1}^m \\ y_{i-1}^m \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{i-1}^m \\ y_{i-1}^m \\ 1 \end{bmatrix}\end{aligned} \tag{4}$$

Then, we just need to compress the affine transform matrix A. There we employ the differential location coding method proposed in [22] with SA mode. For the remnants that are not predicted from f_{i-1}, we use scalar quantization to compress the differences between current location and reference location. The quantization step is $\boldsymbol{Qs}$.

3.4. Global descriptor propagation coding

Since the content of successive video frames have temporal correlation, we observe that global descriptors of two consecutive frames share a large proportion of gauss functions with a minor signal difference between every two gradient vector generated by the same gauss function. Based on this, we proposed an efficient propagation technique to encode global descritpor SCFV. Formally, let G_i denotes the original SCFV vector, and G_i' denote the predictively reconstructed SCFV vector. We evaluate if the similarity between the i-th and (i-1)-th frames satisfy the constrain as follows:

$$\boldsymbol{S}(G_i, G_{i-1}') = \sum s_i^p s_{i-1}^q \boldsymbol{H}(g_i^p, g_{i-1}^q) > \theta \tag{5}$$

Here $\boldsymbol{H}(\cdot,\cdot)$ is a weighted Hamming function introduced in [12]. When the $\boldsymbol{S}(G_i, G_{i-1}')$ exceeds a predefined threshold $\boldsymbol{\theta}$, then assign $G_i' = G_{i-1}'$. So at encoding procedure, we just need to encode 1 bit indicating that the global descriptor of previous frame will be propagated and reused by current frame.

4. EXPERIMENTS

Dataset and Metrics. To evaluate the proposed video descriptors coding framework, both image matching and retrieval tasks were carried out over the Stanford MAR Dataset [15]. The dataset consists of 32 objects including books, CD covers, DVD covers and common objects. These videos were captured with a hand-held mobile phone with different amounts of camera motion, glare, blur, zoom, rotation and perspective changes. Each video is 100 or 200 frames long, recorded at 30 fps with resolution 640 x 480. For each video, a database image (no background noise) is provided. To evaluate the effect of parameters, we select a subset of 4 videos from MAR: Chris Brown Moving (denoted as *chrisbrownmov*); Rascal Flatts (denoted as *rascalflatts*); Multiple Objects including Polish, Wang Book, Monsters Inc (denoted as *multiplev1*); and Multiple Objects including OpenCV, Barry White, Titanic (denoted as *multiplev2*).

At the rest of experiments, all videos in Stanford MAR Dataset are used. To evaluate image matching with localization, we use (i) $N_{inliers}$: the number of feature matches between query video frame and reference image in database using ratio test andRANSAC [3]. (ii) *Jaccard index*: the localization accuracy computed as the overlap ratio between the area of intersection between the ground-truth and the projected quadrilaterals, and the area of their union. To evaluate image retrieval, *mean precise at rank 1st* [18] is adopted, where each video frame is treated as a query. A large-scale retrieval is set up. We use FLICKR1M as the distractor dataset [23], containing 1 million distractor images collected from Flickr.

Implement details. To validate the efficiency of the proposed coding framework, we integrate the proposed coding method into the MPEG CDVS reference software TM 11 [13]. We choose the operating point 4kB as CDVS descriptor size for baseline, because it obtains state-of-the-art performance with moderate size. The quantization parameter for location coding is set $\boldsymbol{Qs}=\mathbf{3}$. The similarity $\boldsymbol{\theta}$

video	chrisbrownmov					rascalflatts					multiplev1					multiplev2				
	Bit rate (Kbps)	Average $N_{inliers}$	Average Jaccard Index	Time Cost(ms)		Bitrate (Kbps)	Average $N_{inliers}$	Average Jaccard Index	Time Cost(ms)		Bitrate (Kbps)	Average $N_{inliers}$	Average Jaccard Index	Time Cost(ms)		Bitrate (Kbps)	Average $N_{inliers}$	Average Jaccard Index	Time Cost(ms)	
				Tracking	Predict				Tracking	Predict				Tracking	Predict				Tracking	Predict
frame-level coding	671.70	79.26	0.97	0.00	0.00	705.01	117.56	0.97	0.00	0.00	702.96	52.92	0.96	0.00	0.00	670.45	64.31	0.97	0.00	0.00
N_{win}=0,N_{refer}=1	155.39	78.73	0.98	0.00	3.75	158.03	116.16	0.98	0.00	4.70	265.76	53.28	0.97	0.00	3.76	321.87	62.43	0.97	0.00	2.81
N_{win}=1,N_{refer}=1	119.90	83.32	0.98	22.80	4.32	127.57	115.07	0.98	34.13	4.81	218.12	53.43	0.96	19.66	4.54	282.53	63.62	0.97	7.31	3.09
N_{win}=3,N_{refer}=1	95.83	87.62	0.97	60.87	4.61	112.90	113.01	0.98	91.06	6.21	190.27	55.19	0.97	60.21	4.22	266.32	66.81	0.97	20.34	3.47
N_{win}=5,N_{refer}=1	84.88	88.47	0.98	91.25	5.37	110.60	112.71	0.97	136.74	5.35	182.78	53.36	0.97	81.74	4.10	262.73	68.00	0.97	31.56	3.26
N_{win}=5,N_{refer}=4	66.48	87.88	0.98	102.54	5.20	80.04	110.29	0.97	155.40	5.66	144.14	51.93	0.97	91.85	4.61	210.57	66.54	0.96	33.73	4.02
N_{win}=5,N_{refer}=8	59.99	87.74	0.97	114.33	12.57	72.17	106.75	0.97	143.39	4.94	140.97	51.45	0.97	99.77	8.49	206.94	65.64	0.97	37.43	4.44

Table 1. Effect of the tracking window size N_{win} and reference number N_{refer} on bitrate, image matching with localization, and the averaged tracking and predict time per frame.

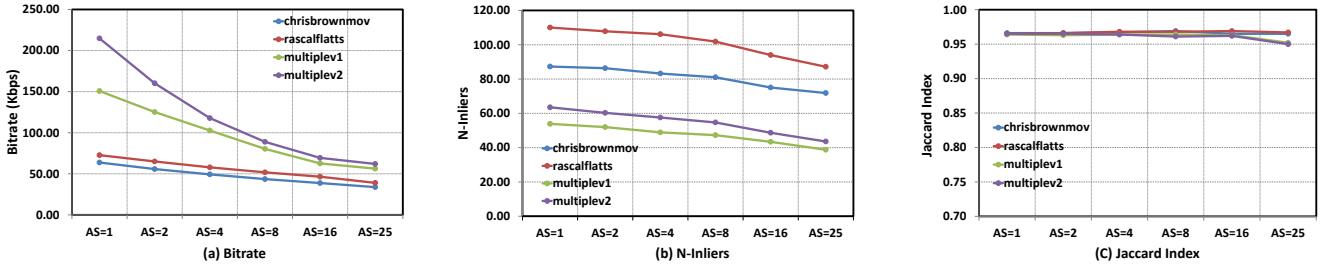

Fig. 4. Effect of the adaptive size AS (from 1 to 25) in ***Skip Mode*** on bitrate, image matching with localization, when $N_{win} = 3$ and $N_{refer} = 8$.

for global descriptors simalarity is trained on MPEG CDVS dataset [23] (including 10,155 matching pairs and 112,175 non-matching pairs), with objective False Positive Rate (FPR) at 1%.

Effect of parameters. In this section, we evaluate the effect of parameters, including N_{win} on keypoints tracking, N_{refer} and AS on compressing local descriptors and location coordinates. As illustrated in Table 1. Firstly, with $N_{refer} = 1$, we observe that the bitrate decreases (50%∼80%) while increasing N_{win} from 0 to 5. The reason is that the number of selected reliable keypoints increases with bigger N_{win} and improves the predictive coding efficiency. Secondly, with $N_{win} = 5$, one can see that the bitrate is further reduced by increasing N_{refer} from 1 to 8, because a growing number of $P - locals$ are predicted from reference local descriptors, resulting in a 3×∼10× bit reduction without performance loss in image matching and localization. As shown in Fig. 2, we evaluate the effect of adaptive size AS (from 1 to 25) in ***Skip Mode*** with $N_{win} = 3$ and $N_{refer} = 8$. It shows that the bitrate is consistently reduced to 30∼60 Kbps as AS increases with little drop in image matching performance. Note that the localization accuracy is stable.

Timing. In Table 1, we measure the average time cost for every frame in both tracking and prediction. We find tracking time grows linearly with window size N_{win}, because of the relatively slow geometric consistency checking. While the prediction time is nearly unchanged when N_{refer} varies, the explanation is that most of local descriptors can find their references in nearest frames.

Performance comparison. For global descriptor coding, the results on whole 32 MAR videos show that the proposed coding scheme obtains an averaged bitrate of 3.57 kbps, compared to 261 kbps with frame-level coding, leading to a 99% bitrate saving. In addition, after performing the reconstructed global descriptors of 3400 video frames as querying in 1M database, the retrieval accuracy in terms of mean precise at rank 1st is 98%, without incurring retrieval accuracy loss. It benefits from the discrimination of SCFV.

To the summary, Table 2 presents the comparison of averaged bitrate and compression ratio between the proposed framework and frame-level coding on the all videos in Stanford MAR dataset. Detailed results are reported for local descriptor coding, location coding and global descriptor coding. It is shown that a high compression ratio of 3.86% is achieved with $N_{win} = 5$, $N_{refer} = 8$ and $AS = 25$. The proposed coding framework offers over 25× fewer bitrate than frame-level coding (i.e., 935.18 kbps vs. 36.07 kbps).

	Frame-level	N_{win}=5 N_{refer}=8, AS=1		N_{win}=5 N_{refer}=8, AS=25	
	Bitrate (Kbps)	Bitrate (Kbps)	Compression ratio	Bitrate (Kbps)	Compression ratio
Local	627.09	134.53	21.45%	18.52	2.95%
Location	47.00	25.98	55.28%	13.99	29.75%
Global	261.09	3.57	1.37%	3.57	1.37%
Total	935.18	164.08	17.55%	36.07	3.86%

Table 2. Comparison of averaged bitrate and compression ratio on the whole Stanford MAR video dataset between the proposed framework and frame-level coding. Detailed results are reported for local descriptor coding, location coding and global descriptor coding.

5. CONCLUSION

In this paper, we propose a high efficient coding framework to address the problem of compressing CDVS descriptors extracted from video sequence. This work is compatible with the MPEG CDVS pipeline and can be easily integrated into CDVS reference software TM 11. Extensive experiments have shown that our approach yielded up to 25× bit rate reduction (less than 40Kbps on average), without any noticeable performance loss in image matching and retrieval tasks. More research works on using rate-distortion optimization techniques will be included in our future work.

6. ACKNOWLEDGMENT

This work was supported by the Chinese Natural Science Foundation under Contracts No. 61271311, No. 61390515, No. 61210005 and by the National Hightech R&D Program of China (863 Program) under Grant No. 2015AA016302.

7. REFERENCES

[1] Bernd Girod, Vijay Chandrasekhar, David M Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S Tsai, and Ramakrishna Vedantham, "Mobile visual search," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 61–76, 2011.

[2] Rongrong Ji, Ling-Yu Duan, Jie Chen, Hongxun Yao, Tiejun Huang, and Wen Gao, "Learning compact visual descriptor for low bit rate mobile landmark search," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, vol. 22, p. 2456.

[3] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.

[5] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477.

[6] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3384–3391.

[7] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3304–3311.

[8] Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua, "Brief: Computing a local binary descriptor very fast," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1281–1298, 2012.

[9] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: an efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.

[10] Stefan Leutenegger, Margarita Chli, and Roland Yves Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2548–2555.

[11] 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 Processing*, vol. 93, no. 8, pp. 2316–2327, 2013.

[12] Jie Lin, L-Y Duan, Yaping Huang, Siwei Luo, Tiejun Huang, and Wen Gao, "Rate-adaptive compact fisher codes for mobile visual search," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 195–198, 2014.

[13] ISO/IEC JTC1/SC29/WG11/N12201, "Cfp for compact descriptors for visual search," 2011.

[14] Ling-Yu Duan, Jie Lin, Jie Chen, Tiejun Huang, and Wen Gao, "Compact descriptors for visual search," *MultiMedia, IEEE*, vol. 21, no. 3, pp. 30–40, 2014.

[15] Mina Makar, Sam S Tsai, Vijay Chandrasekhar, David M Chen, and Bernd Girod, "Interframe coding of canonical patches for mobile augmented reality.," in *ISM*, 2012, pp. 50–57.

[16] Luca Baroffio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi, and Stefano Tubaro, "Coding visual features extracted from video sequences," *Image Processing, IEEE Transactions on*, 2013.

[17] Luca Baroffio, Joao Ascenso, Matteo Cesana, Alessandro Redondi, and Marco Tagliasacchi, "Coding binary local features extracted from video sequences," in *IEEE International Conference on Image Processing*, 2014.

[18] David M Chen, Mina Makar, Andre F Araujo, and Bernd Girod, "Interframe coding of global image signatures for mobile augmented reality," in *Data Compression Conference (DCC), 2014*. IEEE, 2014, pp. 33–42.

[19] ISO/IEC JTC1/SC29/WG11/M22672, "Telecom italia's response to the mpeg cfp for compact descriptors for visual search," 2011.

[20] ISO/IEC JTC1/SC29/WG11/M25929, "Cdvs ce2: local descriptor compression proposal," 2011.

[21] ISO/IEC JTC1/SC29/WG11/M25883, "Cdvs core experiment 3: Stanford/peking/huawei contribution," 2012.

[22] Mina Makar, Vijay Chandrasekhar, S Tsai, David Chen, and Bernd Girod, "Interframe coding of feature descriptors for mobile augmented reality," *Image Processing, IEEE Transactions on*, 2014.

[23] ISO/IEC JTC1/SC29/WG11/N12202, "Evaluation framework for compact descriptors for visual search," 2011.