From Visual Search to Video Compression: A Compact Representation Framework for Video Feature Descriptors

Xiang Zhang†, Siwei Ma†‡, Shiqi Wang†, Shanshe Wang†, Xinfeng Zhang* and Wen Gao†‡

†Institute of Digital Media & Cooperative Medianet Innovation Center, Peking University, Beijing, China

‡Peking University Shenzhen Graduate School, Shenzhen, China

*Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore

†‡{x_zhang, swma, sqwang, wgao}@pku.edu.cn, †sswang@jdl.ac.cn, *xfzhang@ntu.edu.sg

Abstract

Visual feature descriptors have been successfully deployed in a wide range of applications, e.g. visual retrieval and analysis. To transmit these descriptors over bandwidth-limited networks, a high efficiency feature coding technique is highly desired to maximize compression capability and achieve compact feature representations. In this paper, a hybrid visual feature descriptor compression framework is presented and implemented in the encoding and decoding loops of texture videos. In particular, the multiple-hypothesis prediction is employed to effectively remove redundancies originated not only from spatial and temporal similarities, but also from reconstructed video frames. As the ultimate purpose of the transmitted descriptors is retrieval, the rate-accuracy optimization (RAO) technique is proposed to obtain the best tradeoff between the rate and retrieval performance. Such paradigm enables the conventional video stream to achieve high efficient retrieval/analysis with very low bitrate consumption. Moreover, we also demonstrate that texture video compression can also benefit from the additional information provided by the transmitted descriptors, leading to significantly improvement of coding efficiency on top of the high efficiency video coding (HEVC) standard. Extensive simulations have shown that the proposed method can offer significant bitrate reduction in representing both the descriptors and texture video frames, and meanwhile providing desirable retrieval performance.

1 Introduction

With the striking rises in the popularity of hand-held terminals and the exponential increase of images and videos, a variety of mobile visual search and analysis applications emerge with the purpose of linking the virtual and physical worlds, such as landmark recognition, scene recognition and product search. To this end, many effective and robust local descriptors have been developed, e.g., scale-invariant feature transform (SIFT) [1], and speeded up robust features (SURF) [2]. Compression, storage and transmission of such visual features have shown remarkable importance in these applications. Therefore, compact, discriminant and efficient representation techniques of the local feature descriptors are highly desired.

In the literature, prior works mainly focused on the elimination of the intrinsic data dependencies of image descriptors, such as vector quantization [3] locality sensitive hashing (LSH) and [4], Karhunen-Lòeve Transform (KLT) [5]. The recently developed compact descriptors for visual search (CDVS) standard [6] has also been proven to achieve highly efficient retrieval performance with very low bitrate in representing the image feature descriptors. Existing methods on image feature compression can

1068-0314/16 $31.00 © 2016 IEEE
DOI 10.1109/DCC.2016.18

be straightforwardly extended to video applications by coding the descriptors frame by frame, or removing the temporal redundancy with inter-prediction [7, 8]. Particularly, Makar et al. [7] proposed a temporally coherent keypoint detector by forward propagation. Baroffio's work [8] resembled the video coding framework integrating both Intra and Inter frame prediction to compress descriptors.

In another perspective, the visual features have been successfully deployed not only for visual search but also for image (set) [9–11] and video compression [12]. In [9], the feature descriptors were encoded together with a thumbnail image. With the guidance of decoded descriptors, images can be effectively reconstructed by referencing to the similar images in the cloud. In [10, 11], to improve the coding performance, images in the same set that contain strong similarities were dependently coded by feature matching. For video compression, [12] provided a feasible solution, where feature matching techniques was utilized in "merge" and motion vector prediction.

In this work, a hybrid descriptor representation strategy is presented. The distinguished property of our approach is that the descriptors extracted from the video sequence are well represented in the standard video coding framework using very few bits. In this manner, the video coding information can be fully exploited in representing feature descriptors. Specifically, two technical merits, i.e. the multiple-hypothesis prediction and rate-accuracy optimization (RAO), are highlighted in this scheme. Finally, we demonstrate the effectiveness of this paradigm for affine motion estimation in the high-efficiency video coding (HEVC) framework. Experimental results demonstrate that the proposed framework not only provides accurate visual retrieval, but also leads to efficient video coding performance.

2 The Framework of Video Descriptor Representations

The architecture of the proposed scheme is illustrated in Fig. 1. Our work follows the predictive video coding framework, where previously coded features are used to predict the current one, and the residuals after prediction is further compressed by transform, scalar quantization and entropy coding. Particularly, multiple-hypothesis prediction with the devised Intra-frame, Inter-frame and Reconstructed-frame modes, is developed to provide accurate prediction. To optimize this process, the rate-accuracy optimization (RAO) strategy that employs the matching accuracy as the distortion criterion is proposed in best mode selection. Subsequently, the prediction residuals are converted into frequency domain representation with the discrete cosine transform (DCT), followed by scalar quantization and entropy encoding to generate the ultimate feature stream.

Without loss of generality, we demonstrate the framework by compressing the SIFT descriptors, which is one of the most commonly used local features in practical applications. However, it is not limited to SIFT and can be readily extended to other local features. For clarification, some notations are defined as follows. Let $S_j^i = \left(x_j^i, y_j^i, \boldsymbol{\nu}_j^i\right)$ be the j^{th} SIFT feature extracted from i^{th} frame, where x_j^i and y_j^i correspond to the location coordinates, and $\boldsymbol{\nu}_j^i$ denote the 128-dimensional SIFT descriptor vector. In particular, $\widetilde{\boldsymbol{\nu}}_j^i$ and $\widehat{\boldsymbol{\nu}}_j^i$ indicate the reconstructed descriptor and the descriptor extracted from reconstructed frames, respectively.

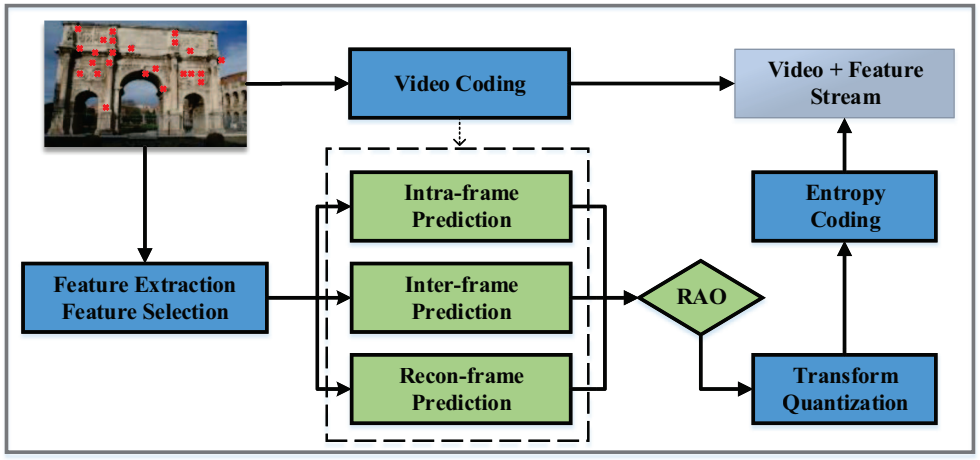


Figure 1: Architecture of the proposed local feature descriptor representation scheme.

3 Technical highlights

3.1 Multiple-Hypothesis Prediction

Multiple-hypothesis prediction is one of the key techniques in video compression standards, such as the H.265/HEVC [13]. In this work, the multiple-hypothesis prediction modes including Intra-frame, Inter-frame and Reconstructed-frame (Recon-frame) predictions, are also employed for providing efficient prediction of video descriptors.

Intra-frame prediction targets at eliminating redundancies caused by non-local similarities in natural images. In this work, the Intra-frame prediction is performed by seeking the best reference descriptor $\widetilde{\boldsymbol{\nu}}_{intra}$ that minimizes the following cost function,

$$\left(\widetilde{\boldsymbol{\nu}}_{intra}, \widetilde{k}\right) = \underset{\widetilde{\boldsymbol{\nu}}_k^i, k\in[0,j)}{\arg\min} \left\|\boldsymbol{\nu}_j^i - \widetilde{\boldsymbol{\nu}}_k^i\right\|_1 + \lambda \cdot R(j-k), \tag{1}$$

where the first term $\left\|\boldsymbol{\nu}_j^i - \widetilde{\boldsymbol{\nu}}_k^i\right\|_1$ denotes the prediction error, measured by the l_1 norm. The second term $R(j-k)$ accounts for the coding rate of the index offset. λ is the Lagrangian multiplier that controls the relative importance between errors and rates. The calculation of the optimal λ will be discussed latter.

The objective of Inter-frame prediction is to remove temporal redundancies between adjacent frames. To efficiently obtain the best matched descriptor in previous frames, we propose to reuse the motion vector (MV) in video coding, which is used in motion compensation from previous frames. The MV is composed of the location offsets (d_x & d_y) and associated reference frame index (d_i). For each being coded local descriptor $S_j^i = \left(x_j^i, y_j^i, \boldsymbol{\nu}_j^i\right)$, the MV (d_x, d_y, d_i) can be derived from the corresponding coding block. In this manner, the search origin is set as $(x_j^i + d_x, y_j^i + d_y)$ in the $(i-d_i)^{th}$ frame. The search set Ψ is restricted to K_Ψ nearest features. Consequently, the optimal Inter-frame prediction descriptor $\widetilde{\boldsymbol{\nu}}_{inter}$ is obtained as follows,

$$\left(\widetilde{\boldsymbol{\nu}}_{inter}, \widetilde{k}\right) = \underset{\widetilde{\boldsymbol{\nu}}_t^{i-d_i}\in\Psi, t\in[0,K_\Psi)}{\arg\min} \left\|\boldsymbol{\nu}_j^i - \widetilde{\boldsymbol{\nu}}_t^{i-d_i}\right\|_1 + \lambda \cdot R(t), \tag{2}$$

where t is the index of the best match in the search set Ψ. Such searching strategy can achieve a good balance between accuracy and computational complexity.

The design philosophy of the third prediction mode is employing the reconstructed frame to extract feature descriptors that can serve as predictors. However, the computational complexity of performing the complete feature extraction in reconstructed frame is considerable for decoder side. Therefore we propose to signal three side-parameters, including the octave o, scale s and orientation θ, in order to skip the keypoint detection process. Both o and s are integers and can be encoded by fixed-length code. However, the orientation is a floating number ranging from $-\pi$ to π, which requires a quantizer for integer conversion. Specifically, let N_θ be the number of orientations after quantization, the optimal $\widetilde{N_\theta}$ can be solved as follows,

$$\widetilde{N_\theta} = \arg\min_{N_\theta} \left(D(N_\theta) + \lambda \cdot R(N_\theta) \right), \tag{3}$$

where $D(N_\theta)$ and $R(N_\theta)$ represent the prediction error and rate respectively, both are the functions of N_θ. Assume the quantized orientation is represented by fixed-length code, then we have $R(N_\theta) = \log_2(N_\theta)$. The relationship between prediction error and N_θ can be fitted as a power function, i.e., $D(N_\theta) = a{N_\theta}^b + c$, by training over large amount of sequences. By incorporating $R(N_\theta)$ and $D(N_\theta)$ into Eqn. (3), the optimal $\widetilde{N_\theta}$ can be calculated as $\widetilde{N_\theta} = \left(\frac{-\lambda}{ab\ln 2}\right)^{1/b}$.

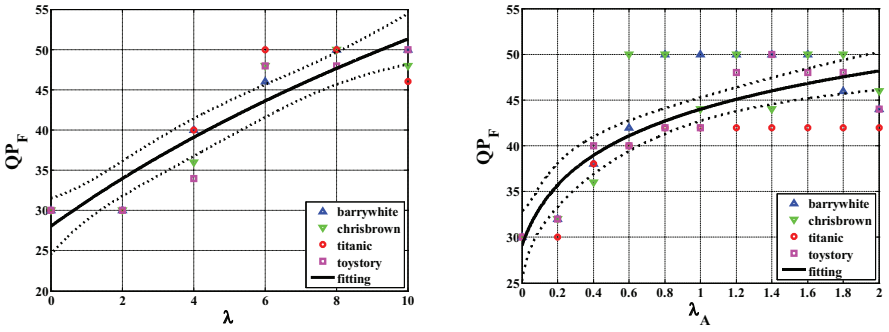
After prediction, the residuals are further compressed by the DCT, scalar quantized and entropy coding sequentially. Here the quantization parameter for descriptor compression is denoted as QP_F to distinguish the one in video coding, which is denoted as QP. For entropy coder, we utilize the context-based adaptive binary arithmetic coding (CABAC) [14] algorithm, which is widely adopted in the HEVC standard.

3.2 Rate-Accuracy Optimization

The rate-distortion optimization (RDO) strategy has been widely adopted in the state-of-the-art video codecs [15–18], for minimizing distortions subject to a constraint bitrate as expressed in Eqn. (1). Generally, the ultimate task of visual descriptors is matching/retrieval. Conventionally applied distortion measures such as sum of absolute difference (SAD) or mean squared error (MSE) cannot reflect the actual degradation in matching performances. To address this issue, we propose a novel rate-accuracy optimization (RAO) approach where the distortion is evaluated in terms of the pair-wise matching accuracy,

$$\min\left(J_A\right), \ \textit{where } J_A = D_A + \lambda_A R, \tag{4}$$

where λ_A indicates the new Lagrangian multiplier. The D_A term quantifies the performance degradation in object matching with the compressed descriptors. However, it is difficult to directly obtain it in the feature encoding process, which inspires us to estimate D_A by the ranking differences in pair-wise matching. Specifically, let $\boldsymbol{\nu}$ and $\widetilde{\boldsymbol{\nu}}$ denote the original and compressed descriptors, and $\mathbb{F}$ represents a collection of descriptors as the matching target. For all descriptors $\boldsymbol{d}^i \in \mathbb{F}$, we get a ranking $\boldsymbol{R_o}$



(a) RDO (b) RAO

Figure 2: The relationship between optimal QP_F and the Lagrangian multipliers. The fitted functions are plotted with 95% confidence bounds. (a) λ in RDO; (b) λ_A in RAO.

in terms of the distance from the original descriptor $\boldsymbol{\nu}$ satisfying that,

$$\underset{i\neq j\in[1,K]}{r_o^i < r_o^j}, \quad s.t. \left\|\boldsymbol{d}^i - \boldsymbol{\nu}\right\|_1 < \left\|\boldsymbol{d}^j - \boldsymbol{\nu}\right\|_1, \tag{5}$$

where $r_o^i, r_o^j \in \boldsymbol{R_o}$ and K is the size of the descriptor collection $\mathbb{F}$. The ranking $\boldsymbol{R_r}$ in terms of the reconstructed descriptor $\widetilde{\boldsymbol{\nu}}$ can be calculated in the similar way,

$$\underset{i\neq j\in[1,K]}{r_r^i < r_r^j}, \quad s.t. \left\|\boldsymbol{d}^i - \widetilde{\boldsymbol{\nu}}\right\|_1 < \left\|\boldsymbol{d}^j - \widetilde{\boldsymbol{\nu}}\right\|_1, \tag{6}$$

where $r_r^i, r_r^j \in \boldsymbol{R_r}$. Then D_A is defined as the differences between the two rankings, which is quantified by the Spearman's rank of correlation coefficient (SROCC),

$$D_A \triangleq 1 - SROCC(\boldsymbol{R_o}, \boldsymbol{R_r}) = \frac{6\sum (r_o^i - r_r^i)^2}{K(K^2-1)}. \tag{7}$$

To derive the Lagrangian multiplier λ and λ_A, a training process is performed to investigate its relationship with QP_F [8,15]. Taking λ_A as instance, the QP_F is varied from 30 to 50 with a step of 2 and the descriptors are compressed using the proposed methods with the corresponding QP_F. As such, the optimal QP_F can be obtained, with which the cost function of Eqn. (4) is minimized. When λ_A ranges from 0 to 2, the corresponding optimal QP_F can be derived in this way. The optimal QP_F in terms of λ and λ_A are plotted in Fig. 2(a) and Fig. 2(b), respectively. Both of them are fitted as a logarithm function $QP_F(\lambda) = a \times \ln(\lambda + b) + c$. With this function, the empirically optimal λ and λ_A can be calculated given QP_F. It is also interesting to note that the RAO curve converges more rapidly than the one from RDO, implying that λ_A is less sensitive to the quantization levels.

To compare the two approaches, all the sequences in database are tested at five different QP_F values. The averaged results are illustrated in Fig. 3, revealing that the proposed RAO method can effectively reduce the coding rate while maintaining almost identical retrieval performance.

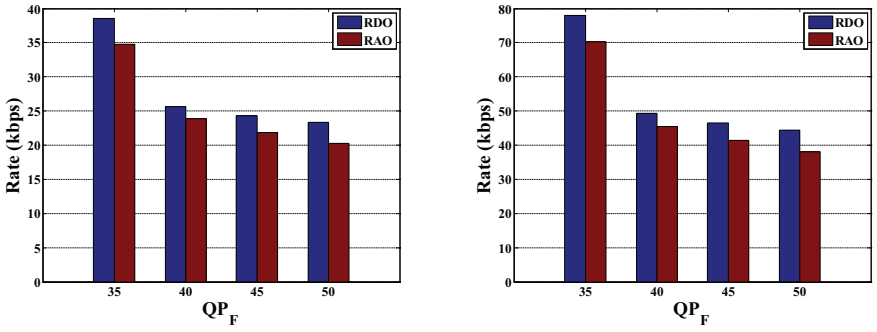


Figure 3: Rate reduction by RAO comparing with the RDO approach, where the maximum number of extracted features in each frame is limited by (a) 100 and (b) 200, respectively.

(a) chrisbrown (b) titanic (c) barrywhite (d) toystory

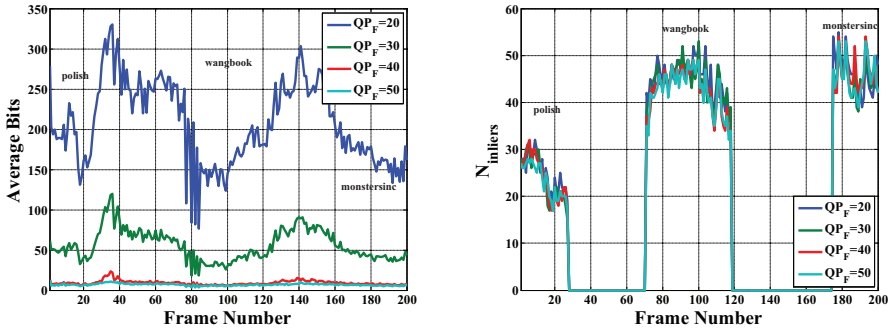
(e) janetjackson (f) privateryan (g) rascalflatts (h) wangbook

Figure 4: Examples of the test sequences in MAR database [19]. First row: sequences with both camera and object motions; Second row: sequences with only camera motions.

4 Experimental Results

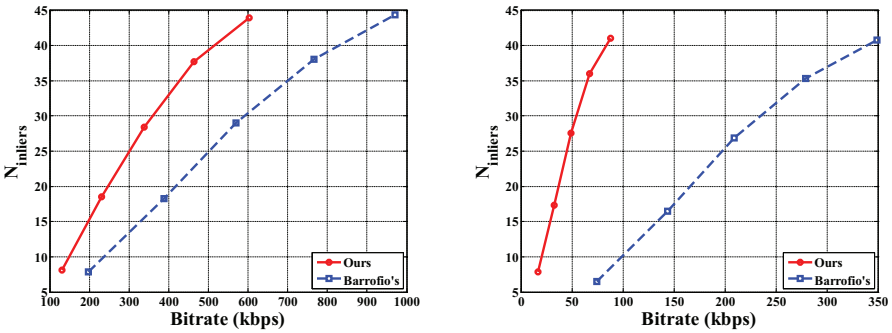
We evaluate the proposed framework of video descriptor compression in the public Stanford streaming mobile augmented reality (MAR) dataset [19]. The dataset consists of 23 different objects of interest. All the video sequences are captured by mobile devices at 30 fps with VGA resolution (640×480), some of which are illustrated in Fig. 4. The HEVC reference software HM-14.0 is used for compressing the video frames with the main profile low-delay P (LDP) configuration. For fair comparison, the pair-wise matching accuracy is adopted as the criterion for performance evaluation. Specifically, the number of the matched descriptor via homograph evaluation [20], i.e. $N_{inliers}$, is obtained for performance evaluation.

Firstly, we evaluate the performances by varying the value of QP_F, which controls the fidelity of descriptors. Fig. 5 gives the results, where the test video contains three interest objects and they show up one after another. One can discern that the bits is dramatically decreased with the increasing QP_F while the matching performance (i.e. $N_{inliers}$) is slightly influenced. For the case of $QP_F = 50$, only 6.85 bits is required for



(a) Average bits (b) $N_{inliers}$

Figure 5: Simulation results with the test sequence that contains three retrieval objects, including polish, wangbook and monstersinc.



(a) $QP_F = 25$ (b) $QP_F = 35$

Figure 6: Performance comparison with the Barrofio's framework [8] under relative high ($QP_F = 25$) and low ($QP_F = 35$) conditions.

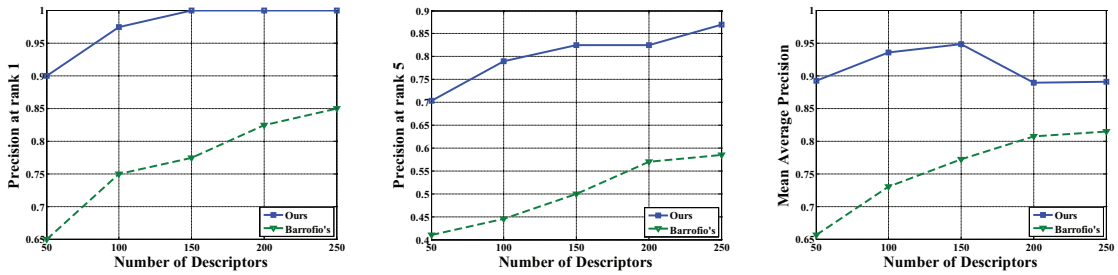
each SIFT descriptor on average, reaching approximately 150:1 compression ratio.

Then we compare our work with Barrofio's scheme [8], in which only Intra- and Inter- predictions and the RDO in mode decision are employed. The experiments are conducted on relatively high ($QP_F = 25$) and low bitrate ($QP_F = 35$) conditions, as shown in Fig. 6(a) and 6(b), respectively. Both of the two sets consist of five operating points that allows different number of extracted features. One can discern that the proposed framework provides remarkable performance gains compared to the Barrofio's, especially for low bitrate cases.

5 Applications

5.1 Landmark Retrieval

In this subsection, we demonstrate the effectiveness of the proposed scheme in the scenario of landmark retrieval. To evaluate the performance, the Rome landmark database [21] is employed, where 10 video sequences of different landmarks are involved as queries. Each query video corresponds to 9 database images on average



(a) Precision at rank 1 (b) Precision at rank 5 (c) mAP

Figure 7: Landmark retrieval performances comparing with the Barrofio's framework [8]. Horizontal axis: the number of compressed descriptors per frame. Vertical axes: (a) precision at rank 1, (b) precision at rank 5, and (c) mean average precision (mAP), respectively.

while the other 10K images serve as distracters.

For each sequence, the frames that contain the target object as well as the corresponding descriptors are compressed using the proposed approach. Retrieval is performed at the receiver side by taking each frame as a query. The similarity criterion between query and target is measured by the number of matched features. The precision at rank k and mean average precision (mAP) are used to evaluate the retrieval performance. The results are illustrated in Fig. 7, which confirm that the proposed scheme outperforms the Barrofio's approach. This further demonstrates that the proposed scheme can be feasibly applied in practical applications.

5.2 Affine Motion Estimation

Local features have the invariance property for camera motion, illumination changing, and viewpoint alternating, etc. In view of this, the compressed feature descriptors are further employed to establish the affine motion model, which covers more complex motions than the traditionally assumed translational moving. The advantages are twofold. First, except the already compressed local features, no additional information is required to transmit; Second, the local features can serve as a faithful source that provides an accurate estimation of the affine motion parameters.

Specifically, the affine motion estimation strategy is implemented on H.265/HEVC platform, where the affine prediction mode is performed for each prediction unit (PU). With the compressed descriptors in video stream, the motion parameters can be straightforwardly derived by feature matching and RANSAC verification. The computed parameters are further refined by gradient descent method for minimizing the differences between the current block and reference block [22]. The affine-merge and affine-skip modes are further developed to allow current PU to derive affine parameters from neighboring blocks for information sharing.

In the experiments, the QP_F is set to be 48 and the maximum number of extracted features in each frame is limited by 100. Table 1 summarizes the BD-rate [23] savings of Y component under LDP configuration. It can be observed that significant bitrate savings can be achieved, even when the coding bits of descriptors are taken

into account. Moreover, the performance gain is highly dependent on the content and video types. For example, the sequence that contains large motions or complex textures will benefit more from the proposed affine motion estimation strategy.

Table 1: Y-component BD-rate reductions via affine motion prediction. Left sequences: both camera and object motions. Right sequences: only camera motions. Two BD-rate indices are applied, i.e. video rate only and total rate including both video and descriptor.

Moving obj. sequences	**BD-rate (Y)**		**Static obj. sequences**	**BD-rate (Y)**	
	video rate	**total rate**		**video rate**	**total rate**
chrisbrown	-14.30%	-12.34%	janetjackson	-4.50%	-2.57%
titanic	-12.95%	-10.04%	privateryan	-6.76%	-1.24%
barrywhite	-7.65%	-4.61%	rascalflatts	-17.82%	-16.41%
toystory	-11.47%	-7.90%	wangbook	-20.69%	-18.45%

6 Conclusion

We proposed a hybrid framework for compact representation of video feature descriptors. The novelty of our approaches lies in efficiently removing the redundancies in video local feature descriptors and optimizing the performance based on the retrieval performance. The superior performance of the proposed scheme was demonstrated by incorporating it into the video compression process, which offered significant rate reduction (approximately 150:1 compression ratio), while maintaining the state-of-the-art matching performance. Applications of such representation strategy for landmark retrieval and video compression were further demonstrated to prove its effectiveness.

Acknowledgment

This work was supported in part by the Major State Basic Research Development Program of China (2015CB351800), in part by the National Science Foundation (No. 61322106 and No. 61571017), and Shenzhen Peacock Plan, which are gratefully acknowledged.

References

[1] D. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV*. Springer, 2006, pp. 404–417.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2161–2168.

[4] C. Yeo, P. Ahammad, and K. Ramchandran, "Rate-efficient visual correspondences using random projections," in *IEEE International Conference on Image Processing*, Oct 2008, pp. 217–220.

[5] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," in *IS&T/SPIE Electronic Imaging*, 2009, pp. 725 710–725 710.

[6] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. A. Reznik, "Mobile visual search: Architectures, technologies, and the emerging mpeg standard," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 86–94, 2011.

[7] M. Makar, V. Chandrasekhar, S. Tsai, D. Chen, and B. Girod, "Interframe coding of feature descriptors for mobile augmented reality," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3352–3367, Aug 2014.

[8] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding visual features extracted from video sequences," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2262–2276, May 2014.

[9] H. Yue, X. Sun, J. Yang, and F. Wu, "Cloud-based image coding for mobile devices–toward thousands to one compression," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 845–857, June 2013.

[10] Z. Shi, X. Sun, and F. Wu, "Photo album compression for cloud storage using local features," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 4, no. 1, pp. 17–28, March 2014.

[11] X. Zhang, Y. Zhang, W. Lin, S. Ma, and W. Gao, "An inter-image redundancy measure for image set compression," in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, May 2015.

[12] X. Zhang, S. Wang, S. Wang, S. Ma, and W. Gao, "Feature-matching based motion prediction for high efficiency video coding in cloud," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015, pp. 1–6.

[13] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[14] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 620–636, July 2003.

[15] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov 1998.

[16] X. Zhang, S. Wang, S. Ma, and W. Gao, "Towards accurate visual information estimation with entropy of primitive," in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, 2015.

[17] S. Ma, X. Zhang, S. Wang, J. Zhang, H. Sun, and W. Gao, "Entropy of primitive: From sparse representation to visual information evaluation," *IEEE Trans. Circuits Syst. Video Technol.*, 2016.

[18] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, April 2012.

[19] M. Makar, S. S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *Int. Journal of Semantic Computing*, vol. 7, no. 01, pp. 5–24, 2013.

[20] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[21] L. Baroffio, "Rome landmark dataset [online]," https://sites.google.com/site/greeneyesprojectpolimi/downloads/datasets/rome-landmark-dataset.

[22] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 497–501, Mar 2000.

[23] G. Bjontegaard, "Calcuation of average PSNR differences between RD-curves," *ITU-T SG16/Q.6 VCEG document VCEG-M33*, 2001.