Joint Feature and Texture Coding: Toward Smart Video Representation via Front-End Intelligence

Siwei Ma, Xiang Zhang, Shiqi Wang, Xinfeng Zhang, *Member, IEEE*, Chuanmin Jia, and Shanshe Wang

***Abstract*— In this paper, we provide a systematical overview and analysis on the joint feature and texture representation framework, which aims to smartly and coherently represent the visual information with the front-end intelligence in the scenario of video big data applications. In particular, we first demonstrate the advantages of the joint compression framework in terms of both reconstruction quality and analysis accuracy. Subsequently, the interactions between visual feature and texture in the compression process are further illustrated. Finally, the future joint coding scheme by incorporating the deep learning features is envisioned, and future challenges toward seamless and unified joint compression are discussed. The joint compression framework, which bridges the gap between visual analysis and signal-level representation, is expected to contribute to a series of applications, such as video surveillance and autonomous driving.**

***Index Terms*— Video compression, feature compression, front-end intelligence.**

I. Introduction

RECENT years have witnessed an explosion of video big data, especially for the surveillance video [1] which is becoming the biggest big data in the digital universe. In particular, according to the prediction from NVIDIA [2], there will be 1 billion cameras deployed in 2020 producing extreme high-volume data in real-time. The large-scale visual data are of paramount significance for social security, smart city and intelligent manufacturing. However, it is usually impractical to rely on manpower only for the utilization of the video big data, especially in the scenario of safeguarding which requires real-time monitoring and rapid response. Recently, the advances of computer vision have substantially promoted the performance of visual analysis and enabled them to be applied in practical application scenarios. As such, in these circumstances, efficient management of the large scale visual data is highly desired.

Along with the advances of hardware processing capabilities, the intelligence has been gradually enabled in front-end cameras. The emergence of intelligent front-end has brought video big data management and representation tremendous opportunities. More specifically, not only the video compression performance can be improved from the increased computational capabilities, but also the visual analysis tasks can be supported at the front-end. Under the constraint of limited bandwidth, the lossy compression of the video data leads to inevitable analysis performance degradation. As such, the exploration of the structural description with higher-level features from the pristine videos can adequately address these issues and enhance the availability of video big data. This motivates the feature compression, which has emerged as an active topic of great significance to both academia and industry. In view of the importance of feature coding, in contrast to the traditional “compress-then-analyze" (CTA) framework as illustrated in Fig. 1, an alternative paradigm “analyze-then-compress” (ATC) was proposed in [3] and [4]. In particular, ATC aims to extract visual features at the front-end, which are subsequently compressed and delivered to the server side for analysis purpose. Due to the fact that the required bitrate in representing the feature is far less than the texture (here in this work, we use the “texture” as an indication of images or video frames), feature compression can greatly facilitate the simultaneous transmission of multiple video streams, especially in the scenarios of surveillance video communication.

Handcrafted features, which can be typically categorized into global and local feature descriptors, have been intensively investigated in the literature. Global features usually serve for a quick search from large-scale datasets, but they may fail in applications such as object matching and localization due to the lack of local information. The local feature descriptors can refine the search as they are usually invariant to the variances caused by camera motion, illumination changing, occlusion and different viewpoints. Local image descriptors such as Scale-Invariant Feature Transform (SIFT) [5] and Speeded Up Robust Features (SURF) [6], were widely adopted in visual analysis tasks due to the robustness in scale-invariant

Manuscript received May 4, 2018; revised August 17, 2018; accepted September 17, 2018. Date of publication October 1, 2018; date of current version October 2, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61571017 and Grant 61632001, in part by the National Basic Research Program of China 973 Program under Grant 2015CB351800, in part by the Top-Notch Young Talents Program of China, High-performance Computing Platform of Peking University, Hong Kong RGC Early Career Scheme under Grant 9048122 (CityU 21211018), in part by the City University of Hong Kong under Grant 7200539/CS, and in part by the China Scholarship Council under Grant 201706010248. This paper was recommended by Associate Editor W. Liu. *(Corresponding author: Shanshe Wang.)*

S. Ma, X. Zhang, C. Jia, and S. Wang are with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: sswang@pku.edu.cn).

S. Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong.

X. Zhang is with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90007 USA.

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

Digital Object Identifier 10.1109/TCSVT.2018.2873102

TABLE I
SUMMARY OF POPULAR IMAGE FEATURE DESCRIPTORS, WHERE THEIR MAJOR CHARACTERISTICS INCLUDING THE DATA TYPE AND FEATURE DIMENSION ARE COMPARED

Feature Descriptors		Data Type	Dimension	Characteristics
Local features	SIFT [5]	Float	128	Gradient based
	SURF [6]	Float	64/128	Integral-image based
	BRIEF [7]	Binary	128/256/512	Intensity difference tests
	ORB [8]	Binary	256	Based on BRIEF
	BRISK [9]	Binary	512	Configurable circular sampling pattern
	USB [10]	Binary	64	Highly discriminative & fast
Global features	BoW [13]	Float	N*	SIFT feature aggregation
	FV [14]	Float		PCA & binary features
	VLAD [15]	Float		Intensity difference tests
	REVV [17]	Float		Sign binarization & LDA

*It should be noted that the dimensions of global features are application-dependent.

representation. To simplify them, algorithms were developed to generate the binarized local feature with high adaptability, high quality and low computational cost. Hence, the generation process of keypoints and feature descriptors, as well as the matching algorithms for binary descriptors were investigated such as BRIEF [7], ORB [8], BRISK [9] and USB [10]. Descriptors binarization offers more flexibility in the sense of computational complexity, such that real-time processing can be feasibility achieved. Recently, numerous algorithms have been proposed to compactly represent these handcrafted features. For example, several algorithms have been proposed to compress SIFT feature using transform coding [11] as well as vector quantization [12]. On the other hand, the global image descriptors statistically summarize the high level image properties, and the most prominent ones include Bag-of-Words (BoW) [13], Fisher Vector (FV) [14] and VLAD [15]. For compact global descriptor representation, efforts have also been devoted to reducing the bits of image signatures, such as tree-structure quantizer [16] for BoW histogram. Regarding VLAD, Chen *et al*. [17] specifically introduced Residual Enhanced Visual Vector (REVV) by reducing the VLAD dimensions with Linear Discriminative Analysis (LDA), which is performed followed by sign binarization. As a summary, the major characteristics of the most popular image feature descriptors are provided in Table I.

For video feature descriptors, the redundancies in temporal domain can be further removed by exploiting statistical dependencies. In [18] and [19], both intra- and inter-feature coding modes are designed to compress SIFT- and BRIEF-like [7] descriptors, and the best mode is selected based on an advanced mode decision strategy. Towards mobile augmented reality application, Markar *et al*. [20] introduced a temporally coherent keypoint detector to perform inter-frame coding of canonical patches. Moreover, for global descriptors, Chen *et al*. [21] Chen and Girod [22] proposed the inter-feature compression scheme for scalable residual based global signatures.

In view of the great importance of compact feature representation, to further enable the inter-operability with the standardized bitstream syntax, MPEG has standardized the "compact descriptors for visual search" (CDVS) [23]–[25] for low bitrate representation of image descriptors. In particular, the CDVS standard adopts the block-based frequency domain Laplacian of Gaussian interest point detector [26]. For local descriptors, a compact SIFT compression scheme with transform followed by ternary scalar quantization is adopted [27]. The location coordinates of the local features are also compressed by representing them with a histogram consisting of a binary histogram map and a histogram counts array [28]. For global descriptors, CDVS adopts the scalable compressed Fisher Vector (SCFV) [29] representation for image retrieval tasks. In particular, the selected SIFT descriptors are aggregated to the FV, and to compress the high dimensional FVs, a subset of Gaussian centroids from the Gaussian Mixture Models (GMM) are selected. For the task of compact descriptors representation for videos, MPEG has also started the standardization of Compact Descriptors for Video Analysis (CDVA) [30]–[32], aiming to provide the inter-operable design and standardize the bitstream to meet the growing demand of video analysis. It is also worth mentioning that the deep learning based features have been adopted in CDVA, due to its superior performance in video analysis tasks.

In this work, we systematically review and analyze the joint feature and texture coding framework, including the design philosophy, advantages and future challenges. In particular, we will start by a brief overview of the joint coding framework and demonstrate its superiority quantitatively and qualitatively. Subsequently, we show that the feature and texture are not necessarily needed to be independently compressed, and the interactions between them can further boost the joint compression performance. In particular, the compressed features can be applied to predict more accurate motion information between neighboring video frames, thus the texture compression performance can be significantly improved. Moreover, the retrieved similar image contents from cloud by matching the decoded features can further be utilized in high quality frame restoration. Furthermore, due to the considerable number of deep learning algorithms which have been repeatedly

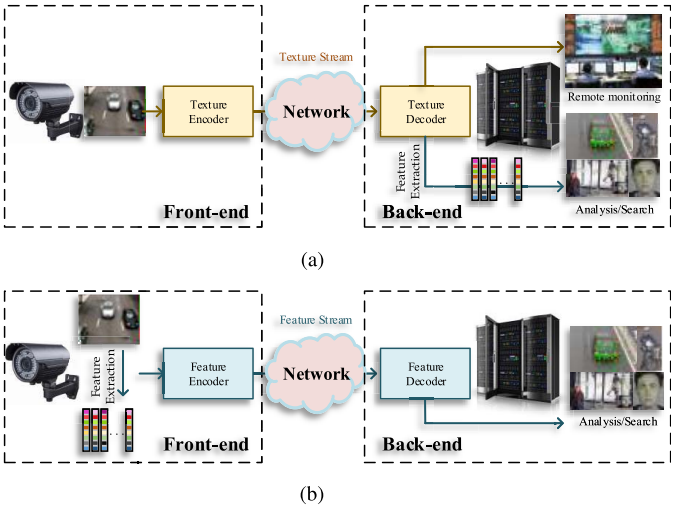


Fig. 1. Two prominent paradigms for visual data management. (a) Compress-then-analyze. (b) Analyze-then-compress.

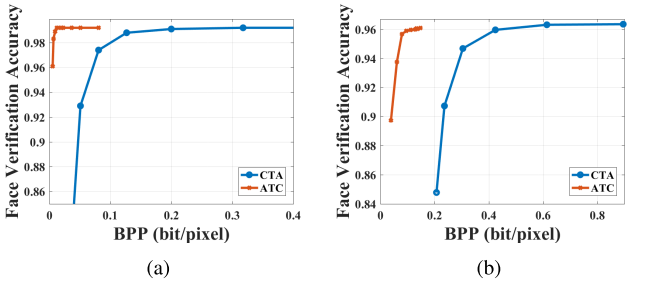


Fig. 2. Face verification performance comparison between ATC and CTA. (a) Facenet [40] on LFW dataset [42]; (b) Deepid [41] on Youtube Faces Database [43].

proved to achieve superior performance in visual analysis tasks, the future work of joint deep learning feature and texture are also thoroughly discussed. It is worth mentioning that in this paper, all the experiments for handcrafted features are conducted with CPUs only, and for deep learning based features GPUs are also adopted.

The remainder of this paper is organized as follows. In Section II, we present the joint feature and texture coding framework, and the advantages as well as superiority of this framework are demonstrated. In Section III, the interactions between feature and texture compression especially for the leverage of the features in improving the texture compression performance are discussed. The joint texture and deep learning based feature compression is envisioned in Section IV and this paper is concluded in Section V.

II. Joint Compression of Feature and Texture

In this section, we introduce the joint compression architecture of feature and texture, which is a hybrid framework that incorporates the advantages of both the CTA and ATC schemes. As shown in Fig. 1, CTA attempts to compress the textures acquired at the front-end, the bitstream of which is subsequently transmitted to the server-end for feature extraction, analysis and viewing [33], [34]. By contrast, ATC first analyzes the captured image/video by extracting visual features at the front-end, which are conveyed to the server side enabling highly light-weight and efficient visual analysis.

Therefore, both CTA and ATC have their own advantages in specific application domains. For example, in the scenario of visual analysis, ATC is superior to CTA in several aspects. First, features are usually much more compact than textures, making the transmission more efficient. As such, in the scenario of low bitrate transmission, the ATC approach is the preferable solution as there exists a lower bound for the CTA paradigm. For example, as analyzed in [4], the source rate is required to be higher than 17.3 kbyte/query for Oxford data set [35] and 2.4 kbyte/query for ZuBud data set [36]. Second, performing analysis with the features from the quality-degraded texture would significantly degrade the analysis performance [37], [38], and ATC overcomes this by extracting the features from the pristine texture. We conducted quantitative comparison between CTA and ATC paradigms in terms of the rate and analysis accuracy. As a demonstration, the experiments are based on the task of face verification. For the CTA paradigm, the state-of-the-art video coding standard High-Efficiency Video Coding (HEVC) [39] is adopted to compress the facial images with the all intra (AI) configuration, where the quantization parameters (QP) are set to be {22, 27, 32, 37, 42, 47, 51} for a large range of bitrate. Then the two deep network models in [40] and [41] are utilized to extract the features for face verification based on the compressed images from different datasets. As for the ATC paradigm, the deep network model firstly applies on the uncompressed images to obtain raw deep features. Then, the extracted deep features are scaled and quantized into integers and then entropy coded. The quantization detail for deep features is as follows,

$$Q_{step} = \frac{2^{\frac{QP-4}{6}}}{s} = 2^{\frac{QP-4}{6}-10}, \overline{C} = floor\left(\frac{C}{Q_{step}}\right), \quad (1)$$

where s indicates the scaling factor (equaling to 2^{10}), C and $\overline{C}$ denote feature coefficients before and after quantization, respectively. All the quantized deep feature coefficients compose of the feature vector for the verification task. In Fig. 2(a), we use Facenet model [40] as feature extractor to obtain the 128-dimension face feature vector. The performance of two paradigms with face verification task on LFW dataset is illustrated, and obviously ATC outperforms CTA with a clear margin especially under low bitrate circumstances. In addition, similar phenomena can also be observed in Fig. 2(b) when evaluating on the Youtube Faces Database with Deepid model [41] to extract the 160-dimension face feature vector. The CTA paradigm has significant performance drop when the bitrate is relatively low. However, the ATC paradigm could achieve comparable performance using much less bandwidth. Finally, ATC does not require the decoding and feature extraction of texture at the sever end, which enables the energy-efficient multimedia systems. However, ATC also has its own limitations, especially when the video texture is required to be further viewed or monitored at the server end, or multiple types of features are needed to be extracted simultaneously.

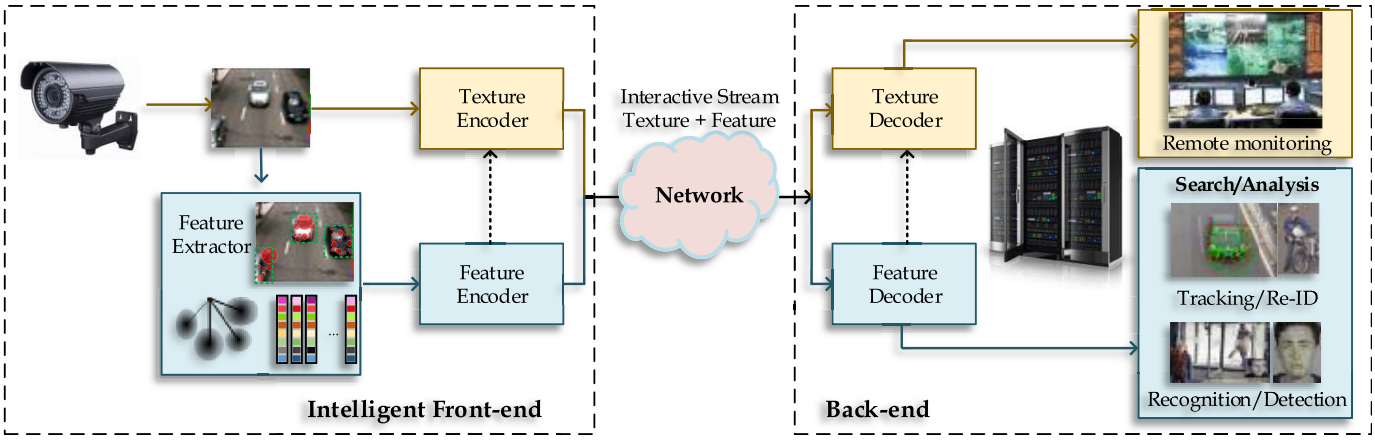


Fig. 3. JFTR paradigm for visual search/analysis applications.

Therefore, the joint feature and texture representation (JFTR) has been proposed to address these problems, and the architecture is illustrated in Fig. 3. In a nutshell, JFTR is a superset of ATC and CTA that compresses and transmits both textures and features [44]–[47]. In analogous to ATC, JFTR extracts and compresses the features at the front-end, such that the features extracted from the pristine image/video can be conveyed and utilized. Moreover, instead of transmitting the compact features only, the compressed texture is also required to be transmitted such that the receiver receives the hybrid texture-plus-feature bitstream. As a hybrid scheme, JFTR overcomes the weaknesses of ATC and CTA by taking advantages of their strengths. In general, ATC and CTA can be regarded as two extreme cases of JFTR. Along with the development of front-end intelligence, in the application scenarios that both analysis and monitoring are required at the server end, the superiority of JFTR makes it to be a desired and promising strategy in the image/video data management.

First, JFTR tackles one major problem of CTA that the visual analysis performance will be degraded when features are extracted from the distorted textures with strong compression artifacts. To demonstrate how would the texture quality influence the analysis performance, we conduct extensive experiments on various analysis-based applications including object matching, object localization and visual retrieval, and the experimental results are shown in Figs. 4&5. To generate textures with varying compression artifacts, the HEVC reference software is utilized for encoding the videos in datasets with different quantization parameters (*i.e.*, QP). Generally speaking, a larger value of QP yields stronger compression artifacts. Here, we utilize the Stanford streaming mobile augmented reality (MAR) dataset [48] for experiments of object matching and localization, where two data categories *Moving* and *Video* are used. The *Moving* category contains videos with camera motion, glare, blur, zoom, rotation and perspective changes, while the *Video* category contains videos with both moving camera and moving objects. For the task of object matching, the performance is evaluated by the number of matched feature pairs between the query (compressed video frame containing the object of interest) and the target

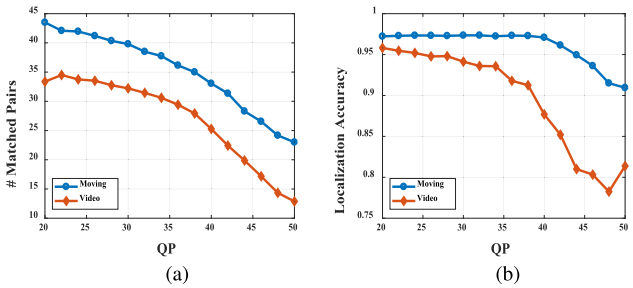


Fig. 4. Object matching and localization performances as a function of QP. Two video categories *Moving* and *Video* in MAR dataset [48] are tested. (a) # matching pairs v.s. QP. (b) Localization accuracy v.s. QP.

(ground-truth image of the object of interest). The object localization accuracy is assessed according to the Jaccard index [49]. As for the visual retrieval task, the Rome landmark dataset (RLD) [50] is utilized, containing 10 Rome landmark videos as queries, and each query has a number of target images while the remaining 10K images are used as distracters. The retrieval performance is evaluated by the precision at rank k and the mean average precision (mAP) indices. In Figs. 4&5, the variations of analysis performance in terms of the quantization parameter (QP) are shown. It is clearly observed that the analysis performance will be significantly degraded at lower bitrate conditions, indicating that under bandwidth limited circumstances CTA approach which transmits the low quality texture is inadequate and unreliable for the purpose of intelligent analysis. For JFTR, the features are extracted from pristine textures, such that the representations of textures can be well preserved and the corresponding high quality features can be extracted for analysis-based applications, leading to significantly improved accuracy.

Second, comparing with CTA, JFTR is more efficient and energy-friendly at the server-end for visual analysis. In [4], the computation and energy complexities of ATC and CTA have been thoroughly compared and analyzed, and it has been shown that CTA is more suitable for deploying at front-end with tight energy constraint, since performing feature

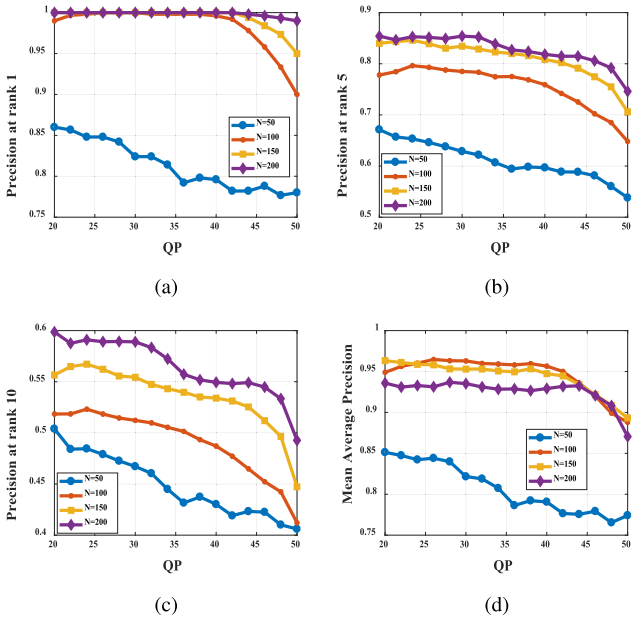


Fig. 5. Retrieval performance as a function of QP. The RLD dataset [50] is used for evaluation. The performance is evaluated with varying maximum number of extracted feature descriptors (*i.e.*, N). (a) Precision at rank 1 v.s. QP. (b) Precision at rank 5 v.s. QP. (c) Precision at rank 10 v.s. QP. (d) mAP v.s. QP.

extraction could consume 2-5 times of energy than performing image encoding. By contrast, at the server end, ATC paradigm is more energy-saving than CTA, as CTA has shifted the complexity of feature extraction from front-end to the server end. With the rapid development regarding the computational capability of the front end cameras, performing feature extraction and compact representation at the front-end is promising and practical. In this sense, JFTR can dramatically reduce the burden of the server and thus improve the quality of experience (QoE) with an instant response. This could be crucial for realtime applications at a resource-constrained platform where a server may receive and process millions of requests simultaneously.

Third, comparing with ATC, JFTR provides extra texture information for human-involved applications such as video monitoring and viewing. ATC attempts to convey compact representation in terms of features, such that significant bitrate reductions can be achieved with the analysis performance being maintained compared with CTA. Nevertheless, the texture information is discarded and ignored. For certain utility that the pixel level information is further required, the texture is of vital importance and should be stored at the server side. As such, the JFTR scheme which takes advantages of both ATC and CTA, could be optimized via joint optimization to simultaneously preserve the texture quality and achieve high efficiency analysis. In other words, the JFTR can achieve similar analysis performance with ATC while maintaining comparable visual quality with CTA, at the expense of conveying both feature and texture bitstream.

In summary, JFTR benefits from both texture and feature compression and enriches smart video representation via front-end intelligence. Though both feature and texture

TABLE III

BITRATE COMPARISON BETWEEN TEXTURE AND FEATURE FOR IMAGES IN LFW DATASET [42].

	Texture Bitrate (bpp)	Feature Bitrate (bpp)	F/T Ratio
OP1	0.4998	0.0229	4.60%
OP2	0.1999	0.015	7.50%
OP3	0.0807	0.0094	11.60%
OP4	0.0367	0.0054	14.70%

bitstreams are required to be transmitted, comparing with textures, the feature data are usually much more compact. We compare the bitrate of texture and feature under four different operation points (OP), where the four OPs range from high bitrate (OP1) to low bitrate (OP4). We report the results on videos and images in Table II and Table III, respectively. For videos, two categories in the MAR dataset [48] are used for evaluation, and the SIFT features are extracted from each video frame and compressed by the method in [47]. Regarding images, the LFW dataset [42] is used and the Facenet [40] is utilized for extracting features. Different QP values are applied to texture and features for different OPs, and for videos the maximum number of extracted features also varies for different OPs. From Table II, one can observe that the feature bitstream only occupies a minority part in the total bitstream, and the ratio of feature bitrate to texture bitrate ranges from 1% to 7% from high bitrate to low bitrate. The ratio in image tests as shown in Table III is relatively higher but still less than 15%. Moreover, the encoded features can further facilitate the texture coding by investigating their interactions [46], [47], [51]. It is also worth noting that the bitrate consumption is highly relevant to the transmission, such that the costs for feature transmission would be relatively lower than texture.

III. FEATURE BASED TEXTURE COMPRESSION

In general, the feature and texture can be compressed individually, or the interactions between them can be further exploited to improve the overall coding performance. In this section, we introduce several techniques that enable the efficient compression of video texture based on the feature representation. The principle behind these techniques is that the feature and texture are redundant to a certain degree, in the sense that features aim to provide informative descriptions of the distinctive characteristics of textures. Moreover, such framework also establishes an interactive work flow for the decoding and utilization of the features and textures. More specifically, the features which are highly compact and frequently visited for retrieval and analysis, can be conveniently accessed without the reference of the texture. By contrast, the texture which occupies a high proportion of the total bitrate, requires the support of the feature information for decoding, such that better coding efficiency can also be ensured.

A. Feature Based Motion Information Derivation

Local feature descriptors, which are required to be highly distinctive and invariant towards camera motion, illumination

TABLE II
BITRATE COMPARISON BETWEEN TEXTURE AND FEATURE FOR VIDEOS IN MAR DATASET [48]

	Moving Category			Video Category		
	Texture Bitrate (Kbps)	Feature Bitrate (Kbps)	F/T Ratio	Texture Bitrate (Kbps)	Feature Bitrate (Kbps)	F/T Ratio
OP1	1870.68	21.64	1.16%	3213.96	22.02	0.69%
OP2	572.89	17.15	2.99%	1057.29	17.63	1.67%
OP3	226.21	12.22	5.40%	427.32	12.75	2.98%
OP4	102.05	6.81	6.67%	194.43	7.21	3.71%

variations and viewpoint altering, can provide more accurate motion predictions among video frames than traditional block-based motion compensation [52]. Previous works [53], [54] have been proposed to utilize local features to derive global motion matrix, which is transmitted to the decoder side for improving inter-frame prediction efficiency. With the JFTR scheme which incorporates features with textures, the global motion information is not required to be conveyed since they are already implied in the compact feature representation. More importantly, the transmitted features can describe more flexible local motions such that the coding efficiency can be further improved.

As presented in [47], the local features extracted from the pristine video sequences are used to facilitate video compression. At the decoder side, the distinctive visual features are first obtained before video decoding, such that affine motion compensation in inter coding can be feasibly supported. In particular, each coding block is assumed to be affine transformed from a patch in a reference frame, and the affine transform can describe the motion with translation, rotation and scaling. Instead of transmitting the affine transform matrix, the matrix can be calculated simultaneously at encoder and decoder sides if there exists matching feature pairs between the current coding block and a reference frame. As shown in Fig. 6, the feature matching is performed between the current block and a corresponding reference frame by ratio test given the decoded features. Then the initial transform matrix $\boldsymbol{T}_0 \in \mathbb{R}^{3\times 3}$ can be obtained with RANSAC algorithm [55] by excluding the outlier matchings. For more accurate motion prediction, $\boldsymbol{T}_0$ is further refined by minimizing the prediction errors as follows,

$$\boldsymbol{T} = \arg\min_{\boldsymbol{T}} \left\{ \sum_i \left[I'(x_i', y_i') - I(x_i, y_i) \right] \right\} \quad \text{s.t.} \quad \begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} = \boldsymbol{T} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \tag{2}$$

where $I(x_i, y_i)$ and $I'(x_i', y_i')$ are pixels in the current block and the corresponding reference frame after affine transformation, respectively. This is solved by gradient descent method, and the differences between initial affine matrix and the refined one $(\boldsymbol{T} - \boldsymbol{T}_0)$ can be transmitted.

Accordingly, the transformed prediction block can be obtained by applying the derived affine matrix to each pixel of current block, and the residual block by subtracting the predicted block from the original block can be represented

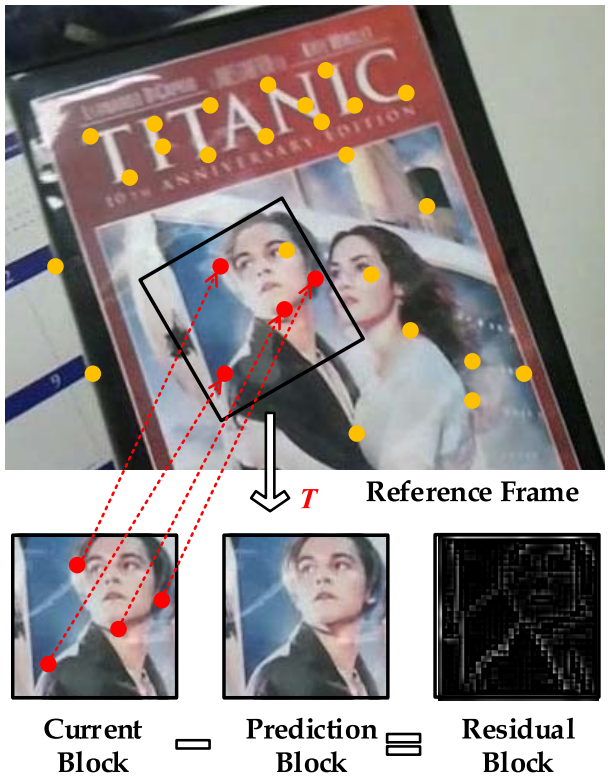


Fig. 6. Affine motion compensated prediction based on feature matching.

in the Discrete Cosine Transform (DCT) domain with less non-zero coefficients, leading to better coding efficiency. The experimental results in [47] have shown significant improvements on video coding efficiency by utilizing the feature matching based inter-frame predictions. The improvements are even more promising especially for video sequences with complex motions, *e.g.*, the videos captured by a moving mobile camera, which contain extensive rotation and scaling motions that conventional translation based block motion cannot efficiently handle. For particular scenario such as mobile captured videos, it has been demonstrated that over 18% bitrate can be saved when applying texture plus feature joint compression compared to the texture compression only. Another benefit shown in [47] is that efficient and accurate visual retrieval can be achieved simultaneously.

An alternative idea is utilizing the derived motion information from features as additional motion vector predictors [56]. In a similar way, the transmitted features can be used for

inter-frame prediction by feature matching. Specifically, for each inter-prediction block, one motion vector can be attained by feature matching between the current block and reference frame. The derived motion vector might be a good predictor of motion vector for current block, therefore it is then added into the motion vector prediction candidates and merge candidates. Spatial and temporal extensions are applied if there is no feature point in current block, where the motion vector can be inherited from larger blocks in the current frame or co-located blocks in neighboring frames. It has been shown that this scheme can bring over 1% bitrate savings, indicating the potential of features in providing more accurate motion information that can be used in video coding.

B. Cloud Based Image and Video Compression

One distinctive property of feature descriptors is that the visually similar patterns and textures can be effectively retrieved based on feature matching. As such, in the scenario of cloud computing that images and videos are stored in the cloud side for further processing and utilizing, the retrieved similar images and videos can play an important role in the redundancy removal to further improve the texture coding performance. In the literature, there are a series of works aiming to improve the quality of reconstructed texture and remove the redundancy within the texture based on hand-crafted features [51], [57]–[61], which validate that the feature based image and video compression is a promising direction for further investigation.

In [58], the cloud based image compression framework was proposed. In particular, the image is described as the corresponding down-sampled version and local feature descriptors such as SIFT. The down-sampled image is then compressed with the traditional image/video codec. Moreover, as the down-sampled version can already provide the rough feature descriptor information, only the residuals between the original SIFT and predicted SIFT from the down-sampled decompressed image are required to be transmitted. As the SIFT residuals are generally sparse and compact, a light version of the image can be obtained and transmitted to the cloud side, achieving much higher compression ratio compared to the traditional image compression standards. In the cloud side, the reconstructed SIFT descriptors can be applied to obtain the correlated and visually similar images in the cloud. In particular, the correlated information can be obtained at the patch level, such that image stitching can be performed with the high quality retrieved patch. The patch stitching is guided based on the distortion between the up-sampled patch from the down-sampled version transmitted to the cloud and the transformed version from the retrieved high quality patch. Due to the external reference information provided, much better compression performance can be achieved when the retrieved correlated image can be obtained. However, one major obstacle of the proposed scheme is that it is difficult to ensure that the correlated images always exist in the decoder side.

The inherent philosophy behind [51] is the external information is very valuable in the image restoration, especially image super-resolution. Inspired by this, a series of works have been proposed for cloud based landmark image super-resolution [59] and image denoising [58]. In particular, for landmark image super-resolution, the difficulty lies in that only the feature descriptors from the up-sampled image are available, such that the Bundled SIFT descriptors [62] are used to obtained the retrieved images. These retrieved images are further aligned based on global registration, and the structure-aware matching criterion was adopted to achieve reliable reconstruction. For cloud based image denoising, the external and internal data cubes are used to filter the image in the 3D transform domain, consisting of a 2D wavelet transform and a 1D Hadamard transform. In order to obtain the similar patches given the noisy version, a graph based matching strategy is adopted to achieve robust patch retrieval.

Inspired by the fact that the feature descriptors exhibit powerful capabilities in finding the high correlated references, the image set compression can be further facilitated by such feature assistant texture coding framework. In [61], the local feature based photo album compression scheme was proposed, which shows much superior performance compared to the HEVC inter/intra coding. In this scheme, the extracted features from each image in an image set are used for two main purposes. First, the similarity of every image pair is estimated by measuring the distance between the matched features of two images. Thus, an efficient prediction structure can be achieved by solving a classic graph problem, where every image in the image set represents a vertex, and the edge between two vertexes represents the similarity of two corresponding images. To maximize the correlations of the image set, the minimum spanning tree (MST) of the graph can be established, leading to more efficient inter-image prediction and higher compression ratio. Second, the features can also be used for guiding the inter-image predictions by a feature-based multi-model geometry transform approach. Traditional block matching based motion search fails to find effective predictions for image set coding, since images in an image set may have very different luminance strengths and viewing perspectives. To tackle this, a three-step prediction strategy including geometric deformation, photometric transformation and block motion compensation was proposed. Geometry deformation aims to compensate global motions caused by the camera perspective difference, where the features are utilized to calculate the deformation matrix by RANSAC algorithm [55]. To increase the prediction accuracy, multiple geometry transformation matrices are obtained by solving a graph-cut problem. After geometric deformation, the photometric transformation is performed in order to reduce the luminance differences between two similar images, where a linear transform model is estimated by minimizing the pixel differences between all matched feature points. After that a block motion compensation is further introduced to eliminate the impact of local shifts. The prediction structure, geometry deformation and photometric transformation parameters are transmitted to the decoder side as additional overhead. In [57], the performance of image set compression is further improved based on the novel inter-image redundancy measure, which was developed for optimizing the prediction structure. In particular, the SIFT feature matching is performed between two

similar images, and the covered area of matched SIFT points and the distance of each pair of matched SIFT descriptors are utilized for calculating the image correlation.

Considering the large-volume near-duplicate videos (NDVs) existing in the cloud, the near-duplicate video compression was proposed in [60]. In this scheme, the GPU based MapReduce framework was employed to NDV retrieval, and the hand-crafted features such as Hessian-Affine detector and SIFT are employed to calculate the similarity. The similar videos are subsequently grouped by establishing the standard minimum-cut problem in graph theory, such that the intra-group similarities can be maximized. After partitioning the videos into several groups among which the intra-group similarity is maximized and inter-group similarity is minimized, the effective compression strategy was subsequently proposed. In particular, one basic video is selected for each group of video based on intra-group similarity, which is encoded independently while the remaining videos are encoded by referencing the corresponding basic video for inter-video prediction. Since the resolutions, luminance and geometric location of videos within a group may be diverse, and the temporally co-located frames in videos with the same group cannot be guaranteed to represent the same visual content, the Inter-Video Reference Index (IVRI) was investigated by analyzing the frame similarity between the basic video and dependent videos to achieve more efficient prediction. Moreover, the Homographic Transformer (HT) and Light Regulator (LR) are deployed for basic videos to produce more accurate reference for inter-video prediction. During the NDV joint coding for dependent videos, there is an additional reference frame to be employed for motion estimation and mode decision besides the traditional temporal intra-video reference frames, and all the other encoding processes remain the same as HEVC encoder. The bitstream of dependently coded video is concatenated with that of the basic video to form the NDV bitstream. To achieve random access capability for joint NDV compression, at the decoder side, the basic video is decoded first, after which the dependency videos can be decoded. The NDV bitstream can also be transcoded into HEVC compliant bitstream by first decoding basic video for inter-video reference frame generation, and subsequently the cascaded transcoding of each dependent video is performed.

IV. Envisioning the Future

Despite significant progress on the joint coding of feature and texture with compact hand-crafted feature representation, the recent advances of deep learning features are also driving the development of many visual analysis tasks. In particular, there is a rather rich literature of deep learning models and features, which are consistently outperforming the hand-crafted features in many vision tasks. Moreover, recent developments on the optimization and acceleration of deep neural networks [63]–[65] have greatly reduced the computational costs, making the deployment of deep neural networks in front-end devices more feasible. With these schemes, more than 10x acceleration and real-time applications of deep neural networks could be achieved with little accuracy degradation. Furthermore, efficient models have been investigated [66] for mobile and embedded vision applications. Energy-efficient pipelines [67] have been proposed to reduce the memory bandwidth of deep neural networks. With these approaches, applying the joint compression schemes in front-end becomes promising in the near future. Here, we will discuss and envision the future joint compression of deep learning features and texture, from the perspectives of deep feature compression, interactions between deep features and video textures in compact representation, as well as the joint bit allocation between video textures and deep features for optimal coding performance.

In contrast to traditional hand-crafted feature descriptors, the compression of deep feature descriptors raises many challenging issues. In particular, deep neural networks such as convolutional neural networks are composed of multiple hierarchical layers (*e.g.*, convolutional and fully connected layers). As such, the identification of the layer to be compressed is meaningful to balance the compact feature size and the representation capability. Moreover, the rapid evolution of deep neural network and the lack of the versatile deep neural network model for various visual analysis tasks also motivate the utilization of the layer with high generalization ability in the joint compression framework. In addition, there exists high redundancy in the representation layers, such that advanced inner- and inter-layer prediction methods are also desired to remove the redundancy. In [68] and [69], the paradigm of collaborative intelligence was proposed to bridge the gap between mobile-only and cloud-only schemes. More specifically, for collaborative intelligence the feature extraction process with deep neural network is split into the front and back ends. For front-end such as the mobile device, the limited computational resources only support the feature extraction at a certain layer, such that these middle features are uploaded to the cloud for further processing and analyses. In [70], the quantization and compression of the features are studied, and the HEVC Range extension with 4:0:0 sampling format [71] is used to compress the quantized features. It is shown that the feature size can be reduced by up to 70% without sacrificing the accuracy. Moreover, the powerful deep learning feature has also been recognized by MPEG when developing the standard of CDVA, and the combinations of deep learning and hand-crafted features are adopted to achieve better video analysis performance. These pioneer studies lay the groundwork for the future investigation of joint deep feature and texture compression. However, the most challenging issue of learning based feature compression and further standardization is how to learn a general feature for various intelligence tasks based on visual contents from different domains. This is still an open question because most of the success we see today is specific task oriented.

Generally speaking, the deep features also imply the low-level texture information, and the recent developments of the deep generative model such as the generative adversarial network (GAN) [72] and variational auto-encoder (VAE) [73] have also demonstrated the powerful ability in generating the texture information given the deep learning features. In [74] and [75], the end-to-end deep learning based image coding framework is comprehensively studied, where the

deep learning features are compressed to recover the texture information. However, most of these works are developed based on the fact that the compressed features targeting at the reconstruction of the textures only, and the visual analysis tasks have not been further taken into account. In [76], based on the observation that the deep learning models used for compression have high similarities with the ones used for inference, the decoding and reconstruction of the texture content is bypassed and two analysis tasks image classification and semantic segmentation are directly performed on the compressed deep learning features. In [77], it is shown that the facial images can be reconstructed from the deep features which are extracted for face recognition, based on which the vulnerabilities of the face recognition based on facial template reconstruction attack are studied.

Given the video texture and features to be transmitted, a natural question is how to allocate the given bit rate budget to each individual such that the final utility can be ensured and optimized. Here, the utility is defined based on the ultimate utilization of the bitstream, including visual viewing/monitoring and analysis. This corresponds to the well established rate-distortion theory in traditional video compression. However, in the joint compression scenario, the distortion term, *i.e.*, the utility involves multiple factors, making the problem rather complicated. In particular, the joint bit allocation process can be formulated as follows,

$$\max_{QP_t, QP_f} \; U(QP_t, QP_f), \quad s.t.\ R_t + R_f \leq R_c, \tag{3}$$

where U denotes the utility function and QP_t and QP_f are the quantization parameters for texture and feature. R_t and R_f are the corresponding bit rate and R_c is the bit rate constraint. The constrained problem can be converted into an unconstrained problem with a Lagrange multiplier λ:

$$\min_{QP_t, QP_f} \; J = -U(QP_t, QP_f) + \lambda \cdot (R_t + R_f) \tag{4}$$

Here, J denotes the joint rate-utility cost. In principle, the utility U can be measured as the combination of the signal level visual quality and visual analysis accuracy, which can also be expressed as functions of their corresponding bit-rates. In [62], the utility function is defined as the Texture-Feature-Quality-Index based on linear combination of the texture quality in terms of MSE and analysis accuracy such as the error rate of face recognition. As such, the optimal bit allocation can be performed by computing the derivation of J to R_t and R_f, respectively. The most challenging issue in the joint bit allocation is the definition of a trusted utility measure as design criteria for optimization. In particular, in video surveillance we may encounter complex application scenarios where the multi-task applications (*e.g.*, face recognition, personal re-identification, car re-identification) are involved. As such, how to reasonably combine these analysis criteria with the texture quality should be further investigated.

Overall, from our perspectives, we believe that the joint compression strategy is as an important direction that deserves further studies and investigations along with the advances of deep learning. First, we believe the learning based especially the deep learning based features will play an important role in the JFTR for intelligence application scenarios. Subsequently, from our point of view, the inverse problem of generating texture from features is also challenging, and it is the key issue in the joint compression loop. This leads to the third direction of joint compression, *i.e.*, how to achieve a good tradeoff between the feature compression and texture compression.

V. Conclusions

In this paper, we have reviewed the advantages of the joint compression of texture and feature, which is an alternative strategy in representing the visual information with the increasing proliferation of intelligent front-end cameras and processing units. We have also reviewed the emerging interaction methods between the texture and feature in the compression process, aiming to further improve the coding efficiency of texture based on the feature information. The core advantages of the joint representation lie in three aspects, including ensured feature quality, energy efficient feature utilization and enhanced coding performance. The potential applications of the joint compression scheme are also discussed and the future development of this joint coding scheme with deep learning based feature extraction is envisioned. As such, here the message we are trying to convey is not that one should abandon the use of video compression nor feature compression. Rather, in some scenarios when both visual monitoring/viewing and analysis are required, the joint scheme is a more powerful alternative that can be practically deployed. It is also expected that further motivations to consider the joint feature and texture coding can be provided towards smart video representation, and great benefits will be brought forward along with the advances of the front-end intelligence. While the video coding and feature coding standardizations will continue to be developed to achieve better representation efficiency within their own application scopes, the standardization of joint feature and texture compression is still in its infancy age. To further enable the interactions and inter-operations, it is also our greater desire to see such standard to be developed and used, especially in applications where both analysis and monitoring might be required to meet the the application requirement.

References

[1] T. Huang, "Surveillance video: The biggest big data," *Comput. Now*, vol. 7, no. 2, pp. 82–91, 2014.

[2] *NVIDIA AI City Challenge*. Accessed: Nov. 1, 2018. [Online]. Available: https://www.aicitychallenge.org/

[3] B. Girod *et al.*, "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011.

[4] A. Redondi, L. Baroffio, L. Bianchi, M. Cesana, and M. Tagliasacchi, "Compress-then-analyze versus analyze-then-compress: What is best in visual sensor networks?" *IEEE Trans. Mobile Comput.*, vol. 15, no. 12, pp. 3000–3013, Dec. 2016.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

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

[7] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2564–2571.

[9] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2548–2555.

[10] S. Zhang, Q. Tian, Q. Huang, W. Gao, and Y. Rui, "USB: Ultrashort binary descriptor for fast visual matching and retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3671–3683, Aug. 2014.

[11] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," *Proc. SPIE*, vol. 7257, p. 725710, Jan. 2009.

[12] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[13] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.

[14] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3384–3391.

[15] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3304–3311.

[16] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *Proc. Data Compress. Conf.*, 2009, pp. 143–152.

[17] D. M. Chen *et al.*, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Process.*, vol. 93, no. 8, pp. 2316–2327, Aug. 2013.

[18] 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 2013.

[19] L. Baroffio, J. Ascenso, M. Cesana, A. Redondi, and M. Tagliasacchi, "Coding binary local features extracted from video sequences," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 2794–2798.

[20] M. Makar, V. Chandrasekhar, S. 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.

[21] D. M. Chen, M. Makar, A. F. Araujo, and B. Girod, "Interframe coding of global image signatures for mobile augmented reality," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2014, pp. 33–42.

[22] D. M. Chen and B. Girod, "A hybrid mobile visual search system with compact global signatures," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1019–1030, Jul. 2015.

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

[24] L.-Y. Duan, T. Huang, and W. Gao, "Overview of the MPEG CDVS standard," in *Proc. Data Compress. Conf. (DCC)*, 2015, pp. 323–332.

[25] L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 179–194, Jan. 2016.

[26] J. Chen, L.-Y. Duan, F. Gao, J. Cai, A. C. Kot, and T. Huang, "A low complexity interest point detector," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 172–176, Feb. 2015.

[27] S. Paschalakis, K. Wnukowicz, M. Bober, A. Mosca, and M. Mattelliano, *CDVS CE2: Local Descriptor Compression Proposal*, Standard ISO/IEC JTC1/SC29/WG11/M25929, 2012.

[28] *CDVS Core Experiment 3: Stanford/Peking/Huawei Contribution*, Standard I. JTC1/SC29/WG11/M25883, 2012.

[29] J. Lin, L.-Y. Duan, Y. Huang, S. Luo, T. Huang, and W. Gao, "Rate-adaptive compact Fisher codes for mobile visual search," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 195–198, Feb. 2014.

[30] *Call for Proposals for Compact Descriptors for Video Analysis (CDVA)—Search and Retrieval*, Standard I. JTC1/SC29/WG11/N15339, 2015.

[31] *Compact Descriptors for Video Analysis: Objectives, Applications and Use Cases*, Standard ISO/IEC JTC1/SC29/WG11/N14507, Valencia, Spain, Mar. 2014.

[32] L.-Y. Duan *et al.* (2017). "Compact descriptors for video analysis: The emerging MPEG standard." [Online]. Available: https://arxiv.org/abs/1704.08141

[33] J. Chao and E. Steinbach, "Preserving SIFT features in JPEG-encoded images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 301–304.

[34] J. Chao, R. Huitl, E. Steinbach, and D. Schroeder, "A novel rate control framework for SIFT/SURF feature preservation in H.264/AVC video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 958–972, Jun. 2014.

[35] *Oxbuildings*. Accessed: Nov. 1, 2018. [Online]. Available: http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

[36] *Zubud*. Accessed: Nov. 1, 2018. [Online]. Available: http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html

[37] A. Zabala and X. Pons, "Effects of lossy compression on remote sensing image classification of forest areas," *Int. J. Appl. Earth Observ. Geoinform.*, vol. 13, no. 1, pp. 43–51, 2011.

[38] A. Zabala and X. Pons, "Impact of lossy compression on mapping crop areas from remote sensing," *Int. J. Remote Sens.*, vol. 34, no. 8, pp. 2796–2813, 2013.

[39] 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, Dec. 2012.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[41] W. Ouyang *et al.*, "DeepID-Net: Deformable deep convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2403–2412.

[42] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Comput. Sci. Dept., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[43] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 529–534.

[44] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Hybrid coding of visual content and local image features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2530–2534.

[45] J. Chao and E. Steinbach, "Keypoint encoding for improved feature extraction from compressed video at low bitrates," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 25–39, Jan. 2016.

[46] X. Zhang, S. Ma, S. Wang, S. Wang, X. Zhang, and W. Gao, "From visual search to video compression: A compact representation framework for video feature descriptors," in *Proc. Data Compress. Conf.*, Mar./Apr. 2016, pp. 407–416.

[47] X. Zhang, S. Ma, S. Wang, X. Zhang, H. Sun, and W. Gao, "A joint compression scheme of video feature descriptors and visual content," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 633–647, Feb. 2017.

[48] 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. J. Semantic Comput.*, vol. 7, no. 1, pp. 5–24, 2013.

[49] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.

[50] GreenEyes. *Rome Landmark Dataset*. Accessed: Nov. 1, 2018. [Online]. Available: https://sites.google.com/site/greeneyesprojectpolimi/downloads/datasets/rome-landmark-dataset

[51] 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, Jun. 2013.

[52] Y.-H. Fok and O. C. Au, "An improved fast feature-based block motion estimation," in *Proc. 1st Int. Conf. Image Process.*, vol. 3, 1994, pp. 741–745.

[53] S.-K. Kim, S.-J. Kang, T.-S. Wang, and S.-J. Ko, "Feature point classification based global motion estimation for video stabilization," *IEEE Trans. Consum. Electron.*, vol. 59, no. 1, pp. 267–272, Feb. 2013.

[54] M. Tok, A. Glantz, A. Krutz, and T. Sikora, "Feature-based global motion estimation using the helmholtz principle," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1561–1564.

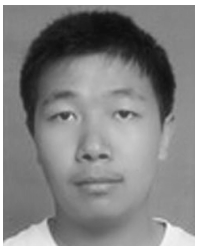
[55] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[56] X. Zhang, S. Wang, S. Wang, S. Ma, and W. Gao, "Feature-matching based motion prediction for High Efficiency Video Coding in cloud," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jun./Jul. 2015, pp. 1–6.

[57] X. Zhang, Y. Zhang, W. Lin, S. Ma, and W. Gao, "An inter-image redundancy measure for image set compression," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 1274–1277.

[58] H. Yue, X. Sun, J. Yang, and F. Wu, "Image denoising by exploring external and internal correlations," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1967–1982, Jun. 2015.
[59] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving Web images," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4865–4878, Dec. 2013.
[60] H. Wang, T. Tian, M. Ma, and J. Wu, "Joint compression of near-duplicate videos," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 908–920, May 2017.
[61] Z. Shi, X. Sun, and F. Wu, "Photo album compression for cloud storage using local features," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 4, no. 1, pp. 17–28, Mar. 2014.
[62] L. Dai, X. Sun, F. Wu, and N. Yu, "Large scale image retrieval with visual groups," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 2582–2586.
[63] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *Fiber*, vol. 56, no. 4, pp. 3–7, 2016.
[64] S. Han *et al.*, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. ACM/IEEE Int. Symp. Comput. Archit.*, Jun. 2016, pp. 243–254.
[65] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 525–542.
[66] A. G. Howard *et al.* (2017). "MobileNets: Efficient convolutional neural networks for mobile vision applications." [Online]. Available: https://arxiv.org/abs/1704.04861
[67] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.
[68] Y. Kang *et al.*, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," in *Proc. Int. Conf. Architectural Support Program. Lang. Oper. Syst.*, 2017, pp. 615–629.
[69] A. E. Eshratifar, M. S. Abrishami, and M. Pedram. (2018). "JointDNN: An efficient training and inference engine for intelligent mobile cloud computing services." [Online]. Available: https://arxiv.org/abs/1801.08618
[70] H. Choi and I. V. Bajic. (2018). "Deep feature compression for collaborative object detection." [Online]. Available: https://arxiv.org/abs/1802.03931
[71] D. Flynn *et al.*, "Overview of the range extensions for the HEVC standard: Tools, profiles, and performance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 4–19, Jan. 2016.
[72] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[73] Y. Pu *et al.*, "Variational autoencoder for deep learning of images, labels and captions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2352–2360.
[74] J. Ballé, V. Laparra, and E. P. Simoncelli. (2016). "End-to-end optimized image compression." [Online]. Available: https://arxiv.org/abs/1611.01704
[75] O. Rippel and L. Bourdev. (2017). "Real-time adaptive image compression." [Online]. Available: https://arxiv.org/abs/1705.05823
[76] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool. (2018). "Towards image understanding from deep compression without decoding." [Online]. Available: https://arxiv.org/abs/1803.06131
[77] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain. (2017). "Face image reconstruction from deep templates." [Online]. Available: https://arxiv.org/abs/1703.00832

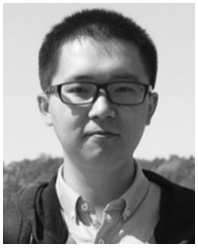
Siwei Ma received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He held a post-doctoral position at the University of Southern California, Los Angeles, CA, USA, from 2005 to 2007. He joined the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing, where he is currently a Professor. He has authored over 100 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission.



Xiang Zhang received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently pursuing the Ph.D. degree with Peking University. His research interests include image/video quality assessment, video compression, and visual retrieval.

Shiqi Wang received the B.S. degree in computer science from the Harbin Institute of Technology in 2008 and the Ph.D. degree in computer application technology from Peking University in 2014. From 2014 to 2016, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. From 2016 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. He has submitted over 40 technical proposals to ISO/MPEG, ITU-T, and AVS standards. His research interests include image/video compression, analysis, and quality assessment.

Xinfeng Zhang (M'16) received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From 2014 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently a Post-Doctoral Fellow with the School of Electrical Engineering System, University of Southern California, Los Angeles, CA, USA. His research interests include image and video processing, and image and video compression.



Chuanmin Jia received the B.Sc. degree in computer science and technology from the Beijing University of Posts and Telecommunications, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Peking University, Beijing, China. From 2017 to 2018, he was a Visiting Student with the Video Lab, New York University, NY, USA. His research interests include video coding, machine learning, and light field image compression.

Shanshe Wang received the B.S. degree from the Department of Mathematics, Heilongjiang University, Harbin, China, in 2004, the M.S. degree in computer software and theory from Northeast Petroleum University, Daqing, China, in 2010, and the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, in 2014. He currently holds a post-doctoral position at Peking University. His current research interests include video compression, and image and video quality assessment.