Feature Article

Compact Descriptors for Video Analysis: The Emerging MPEG Standard

Ling-Yu Duan
Peking University, Peng Cheng Laboratory

Vijay Chandrasekhar
Institute for Infocomm Research, A*STAR

Shiqi Wang
City University of Hong Kong

Yihang Lou
Peking University

Jie Lin
Institute for Infocomm Research, A*STAR

Yan Bai
Peking University

Tiejun Huang
Peking University, Peng Cheng Laboratory

Alex Chichung Kot
Nanyang Technological University

Wen Gao
Peking University, Peng Cheng Laboratory

***Abstract*—This paper provides an overview of the on-going compact descriptors for video analysis standard (CDVA) from the ISO/IEC moving pictures experts group (MPEG). MPEG-CDVA targets at defining a standardized bitstream syntax to enable interoperability in the context of video analysis applications. During the developments of MPEG-CDVA, a series of techniques aiming to reduce the descriptor size and improve the video representation ability have been proposed. This paper describes the new standard that is being developed and reports the performance of these key technical contributions.**

Digital Object Identifier 10.1109/MMUL.2018.2873844
Date of publication 25 October 2018; date of current version 12 June 2019.

Over the past decade, there has been an exponential increase in the demand for video analysis, which refers to the capability of automatically analyzing the video content for event detection, visual search, tracking, classification, etc. Generally speaking, a variety of applications

Published by the IEEE Computer Society

can benefit from the automatic video analysis, including mobile augmented reality (MAR), automotive, smart city, media entertainment, etc. For instance, MAR requires object recognition and tracking in real-time for accurate virtual object registration. With respect to automotive, robust object detection and recognition are highly desirable for warning the collision and cross traffic. The increasing proliferation of surveillance systems is also driving the developments of object detection, classification, and visual search technologies. Moreover, a series of new challenges have been brought forward in media entertainment, such as interactive advertising, video indexing and near-duplicate detection, which all rely on robust and efficient video analysis algorithms.

For the deployment of video analysis functionalities in real application scenarios, a unique set of challenges are presented. Basically, it is the central server that performs automatic video analysis tasks, such that efficient transmission of the visual data via a bandwidth constrained network is highly desired. The straightforward way is to encode the video sequences and transmit the compressed visual data over the networks. As such, features can be extracted from the decoded videos for video analysis purpose. However, this may create high- volume data due to the pixel level representation of the video texture. One could imagine that 470 000 closed-circuit television (CCTV) cameras for video acquisition are deployed in Beijing, China. Assuming that for each video 2.5 Mb/s bandwidth[1] is required to ensure that they can be simultaneously uploaded to the server side for analysis, in total 1.2-TB/s video data are transmitted on the internet highway for security and safety applications. Due to the massive CCTV camera deployment in the city, it is urgently required to investigate ways to handle the large-scale video data.

As video analysis is directly performed based on extracted features instead of the texture, shifting the feature extraction and representation into the camera-integrated module is highly desirable, which directly supports the acquisition of features at the client side. As such, compact feature descriptors instead of compressed texture data can be delivered, which can completely satisfy the requirements of video analysis. Therefore, developing effective and efficient compact feature descriptor representation techniques with low complexity and memory cost is the key to such "analyze then compress" infrastructure. Moreover, the interoperability should also be maintained to ensure that feature descriptors extracted by any devices and transmitted in any network environments are fully operable at the server end. The compact descriptors for visual search (CDVS) standard[1,2] developed by motion picture experts group (MPEG), standardizes the descriptor bitstream syntax and the corresponding extraction operations of still images to ensure interoperability for visual search applications. It has been proven to achieve high efficiency and low latency mobile visual search,[2] and an order of magnitude data reduction is realized by only sending the extracted feature descriptors to the remote server.

However, the straightforward encoding of CDVS descriptors extracted frame by frame from video sequences cannot fulfill the applications of video analysis. For example, as suggested by CDVS, the descriptor length for each frame is 4 K, and for a typical 30-fps video, the feature bit rate is approximately to be 1 Mb/s. Obviously, this may lead to excessive consumption of storage and band- width. Unlike still images, the video combines a sequence of high correlated frames to form a moving scene. To fill the gap between the existing MPEG technologies and the emerging requirements of video feature descriptor compression, a Call for Proposals (CfP) on compact descriptors for video analysis (CDVA)[3] was issued in 2015 by MPEG, targeting at enabling efficient and interoperable design of advanced tools to meet the growing demand of video analysis. It is also envisioned that CDVA can achieve significant savings in memory size and bandwidth resources, and meanwhile provide hardware-friendly support for the deployment of CDVA at the application level. As such, the aforementioned video analysis applications such as MAR, automotive, surveillance and media entertainment can be flexibly supported by CDVA,[4] as illustrated in Figure 1.

In Figure 2, the framework of CDVA is demonstrated, which is comprised of keyframe/shot detection, video descriptors extraction, encoding, transmission, decoding and video analysis against a large scale database. During the development of CDVA, a series of techniques have been developed for these modules. The key technical contributions of CDVA are reviewed in this

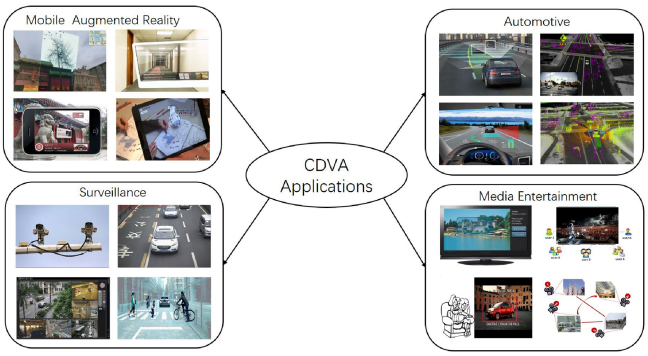

Figure 1. Potential application scenarios of CDVA.

paper, including the video structure, advanced feature representation, and video retrieval and matching pipeline. Subsequently, the developments of the emerging CDVA standard are discussed, and the performance of the key techniques is demonstrated. Finally, we discuss the relationship between CDVS and CDVA and look into the future developments of CDVA.

MPEG CDVS STANDARD

MPEG-CDVS provides the standardized description of feature descriptors and the descriptor extraction process for efficient and interoperable still image search applications. Basically, CDVS can serve as the frame level video feature description, which inspires the inheritance of CDVS features in the CDVA exploration. This section discusses the compact descriptors specified in CDVS, which are capable of adapting the network bandwidth fluctuations for the support of scalability with the predefined descriptor lengths: 512 B, 1 K, 2 K, 4 K, 8 K, and 16K.

Compact Local Feature Descriptor

The extraction of local feature descriptors is required to be completed in a low complexity and memory cost way. Obviously, this is much more desirable for videos. The CDVS standard adopts the Laplacian of Gaussian interest point detector. The low-degree polynomial (ALP) approach is employed to compute the local response after Laplacian of Gaussian filtering. Subsequently, a relevance measure is defined to select a subset of feature descriptors, which is statistically learned based on several characteristics of local features including scale, peak response of the LoG, distance to image centre, etc. Handcrafted SIFT descriptor is adopted in CDVS as the local feature descriptors, and a compact SIFT compression scheme achieved by transform followed with ternary scalar quantization is developed to reduce the feature size. This scheme is of low-complexity and hardware favorable due to fast processing (transform, quantization, and distance calculation). In addition to the local descriptors, location coordinates of these descriptors are also compressed for transmission. In CDVS, the location coordinates are represented as a histogram consisting of a binary histogram map and a histogram counts array. The histogram map and counts array are coded separately by a simple arithmetic coder and a sum context-based arithmetic coder.

Local Feature Descriptor Aggregation

CDVS adopts the scalable compressed Fisher vector (SCFV) representation for mobile image retrieval. In particular, the selected SIFT descriptors are aggregated to the Fisher vector (FV) by assigning each descriptor to multiple Gaussians in a soft assignment manner. To compress the high dimensional FVs, a subset of Gaussian components in the Gaussian Mixture Model are selected based on their rankings in terms of the standard deviation of each subvector. The number of selected Gaussian functions is dependent on the available coding bits, such that descriptor

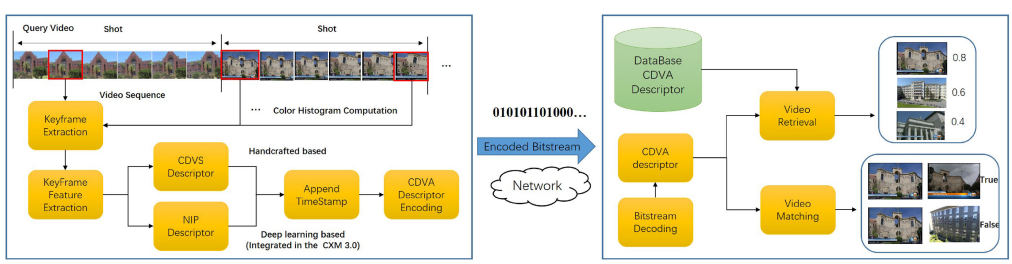

Figure 2. Illustration of the CDVA framework.

scalability is achieved to adapt to the available bit budget. Finally, the one-bit scalar quantizer is applied to support fast comparison with Hamming distance.

KEY TECHNOLOGIES IN CDVA

Driven by the success of MPEG-CDVS, which provides a fundamental groundwork for the development of CDVA, a series of technologies have been brought forward. In CDVA, the key contributions can be categorized into the video structure, video feature description, and video analysis pipeline. The CDVA framework specifies how the video is structured and organized for feature extraction, where key frame detection and interfeature prediction methods are presented. Subsequently, the deep learning based feature representation is reviewed, and the design philosophy and compression methods of the deep learning models are discussed. Finally, the video analysis pipeline that serves as the server side processing module is introduced.

Video Structure

The video is composed of a series of highly correlated frames, such that extracting the feature descriptors for each individual frame may be redundant and lead to unnecessary computational consumptions. In view of this, a straightforward way is to perform key frame detection, following which only feature descriptors of the keyframes are extracted. In the paper by Balestri *et al.*,[5] the global descriptor SCFV in CDVS is employed to compare the distance between the current frame and the previous one. In particular, if the distance is lower than a given threshold, indicating that it is not necessary to preserve the current frame for feature extraction, the current frame is dropped. However, one drawback of this method is that for each frame the SCFV should be extracted, which brings additional computational complexity. In the paper by Balestri *et al.*,[6] the color histogram instead of the CDVS descriptors is employed for the frame level distance comparison. As such, the SCFV descriptors in nonkey frames do not need to be extracted. Due to the advantage of this scheme, it has been adopted into the CDVA experimentation model (CXM) 0.2.[7] Bailer[8] proposed to modify the segment produced by the color histogram. In particular, for each segment, the medoid frame of each segment is selected, and all frames within this segment that have lower similarity in terms of SCVF than a given threshold are further chosen for feature extraction.

The keyframe-based feature representation has effectively removed the video temporal redundancy, resulting in low bitrate query descriptor transmission. However, this strategy has largely ignored the intermediate information between two key-frames. In the paper by Huang *et al.*,[9] it is interesting to observe that densely sampled frames can bring better video matching and retrieval performance at the expense of increased descriptor size. In order to achieve a good balance between the feature bitrate and video analysis performance, the interprediction techniques for local and global descriptors of CDVS have been proposed.[9] Specifically, in the paper by Huang *et al.*,[9] the intermediate frames between two keyframes are denoted as the predictive frame (P- frame). In P-frame, the local descriptor is predicted by the multiple reference frame prediction. For those local descriptors which cannot find corresponding references, they are directly written into the bit-stream. For global descriptors in P-frame, for the component selected from both current and previous frames, the binarized subvector is copied from the corresponding one in the previous frame to save coding bits. In the paper by Bailer *et al.*,[8] it is further demonstrated that more than 50% compression rate reduction can be achieved by applying lossy compression of local descriptors, without significant influence on the matching performance. Moreover, it is demonstrated that the global difference descriptors can be efficiently coded using adaptive binary arithmetic coding as well.

Deep Learning Based Video Representation

Recently, due to the remarkable success of deep learning, numerous approaches have been presented to employ the convolutional neural networks (CNNs) to extract deep learning features for image retrieval.[10,11] In the development of CDVA, the nested invariance pooling (NIP) has been proposed to obtain the discriminative deep invariant descriptors,

and significant video analysis performance improvement over traditional handcrafted features has been observed. It is also worth mentioning that the NIP descriptor has been adopted into the CDVA standard, and the corresponding extraction module is integrated into the reference software. In this section, we will review the development of deep learning features in CDVA from the perspectives of deep learning based feature extraction, network compression, feature binarization, and the combination of deep learning based feature descriptors with handcrafted ones.

Deep Learning Based Feature Extraction Robust video retrieval requires the features to be scale, rotation, and translation invariant. The CNN models incorporate the local translation invariance by a succession of convolution and pooling operations. In order to further encode the rotation and scale invariance into CNN, motivated by the invariance theory, the NIP was proposed to represent each frame with a global feature vector.[12] In particular, the invariance theory provides a mathematically proven strategy to obtain invariant representations with the CNNs. This inspires the improvement on the geometric invariance of deep learning features based on the pooling operations of the intermediate feature maps in a nested way. Specifically, given an input frame, it can be rotated with R times, and for each time the *pool5* feature maps (W H C) is extracted. Here, W and H denote the width and height of the map and C is the number of feature channels. Based on the feature map, the multiscale uniform region of interest (ROI) sampling is performed, resulting in the five-dimensional (5-D) feature reforestation with dimension (R S W' H' C). Here, S is the number of sampled ROIs in multiscale region sampling. Subsequently, NIP performs a nested pooling over translations (W' H'), scales (S), and finally rotations (R). Therefore, a C-dimensional global CNN feature descriptor can be generated. The performance of NIP descriptors can be further boosted by the PCA whitening.[10,11] To evaluate the similarity between two NIP feature descriptors, the cosine similarity function is adopted.

Network Compression CNN models contain millions of neurons, which cost hundreds of MBs for storage. This creates great difficulties in video analysis, especially when the CNN models are deployed at the client side for feature extraction in the "analyze then compress" framework. Therefore, an efficient compression model of the neural network is urgently required for the development of CDVA. In the paper by Bai *et al.*,[13] both scalar and vector quantization techniques using the Lloyd-Max algorithm are applied to compress the NIP model. The quantized coefficients are further coded with Huffman coding. Moreover, the model is further pruned to reduce the model size by dropping the convolutional layers. It is shown that the compressed models that have two orders of magnitude smaller than the uncompressed models lead to a negligible loss in video analysis.

Feature Descriptor Compression The deep learning based feature descriptor generated from NIP is usually in float-point, which is not efficient for the subsequent feature comparison process. As hamming distance can facilitate effective retrieval especially for large video collections, the NIP feature binarization has been proposed for compact feature representation.[12] In particular, the one-bit scalar quantizer is applied to simply binarize the NIP descriptor. As such, much less memory footprint and runtime cost can be achieved with a marginally degraded performance loss.

Combination of Deep Learning Based and Hand-Crafted Features Furthermore, in the paper by Lou *et al.*,[12] it is also revealed that there are some complementary effects between CDVS handcrafted and deep learning based features for video analysis. In particular, the deep learning based features are extracted by taking the whole frame into account while CDVS handcrafted descriptors sparsely sample the interest points. Moreover, the handcrafted features work relatively better in rich textured blobs, while deep learning based features are more efficient in aggregating deeper and richer features for global salient regions. Therefore, the combination of deep learning based features and CDVS handcrafted features has been further investigated in the CDVA framework,[12] as shown in Figure 3. Interestingly, it is validated that the combination strategy achieves promising

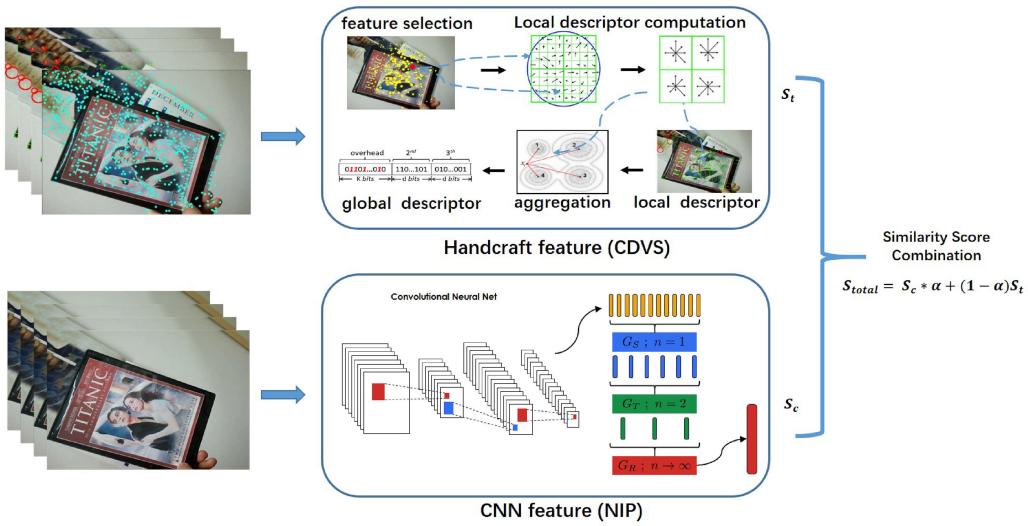

Figure 3. Combination of handcrafted and deep learning based feature descriptors.

performance and outperforms either deep learning based or CDVS handcrafted features.

EMERGING CDVA STANDARD

Video Analysis Pipeline

The compact description of videos enables two typical tasks in video analysis, including video matching and retrieval. In particular, video matching aims at determining if a pair of videos shares the object or scene with similar content, and video retrieval performs searching for videos containing similar segment as the one in the query video.

Video Matching Given the CDVA descriptors of the keyframes in the video pair, pairwise matching can be achieved by comparing them in a coarse to fine strategy. Specifically, each keyframe in one video is first compared with all of the keyframes in the other video in terms of the global feature similarity. If the similarity is larger than the threshold, implying that there is a possible match between the two frames, the local descriptor comparison can be further performed with the geometric consistency checking. The keyframe-level similarity is subsequently calculated by the multiplication of matching scores of the global and local descriptors. Finally, we can obtain the video-level similarity by selecting the largest matching score among all keyframe-level similarities.

Another criterion in video matching is the temporal localization, which locates the video segment containing similar items of interest based on the recorded timestamps. In the paper by Balestri *et al.*,[14] a shot level localization scheme was adopted into CXM1.0. In particular, a shot is detected to be the group of consecutive keyframes whose distance to the first keyframe of this shot is smaller than a certain threshold in terms of the color histogram comparison. If the keyframe-level similarity is larger than a threshold, the shot that contains the keyframe is regarded as the matching interval. Multiple matching intervals can also be concatenated together to obtain the final interval for localization.

Video Retrieval In contrast to video matching, video retrieval is performed in a one-to-N manner, implying that the videos in the database are all visited and the top ones with higher matching scores are selected. In particular, the key-frame level matching with global descriptors is performed to extract the top K_g candidate keyframes in the database. Subsequently, these keyframes are further examined by local descriptor matching, and the keyframe candidate dataset is further shrunk to K_l according to the rankings in terms of the combination of global and local similarities. These keyframes are reorganized into videos, which are finally ranked by the video level similarity following the principle in video matching pipeline.

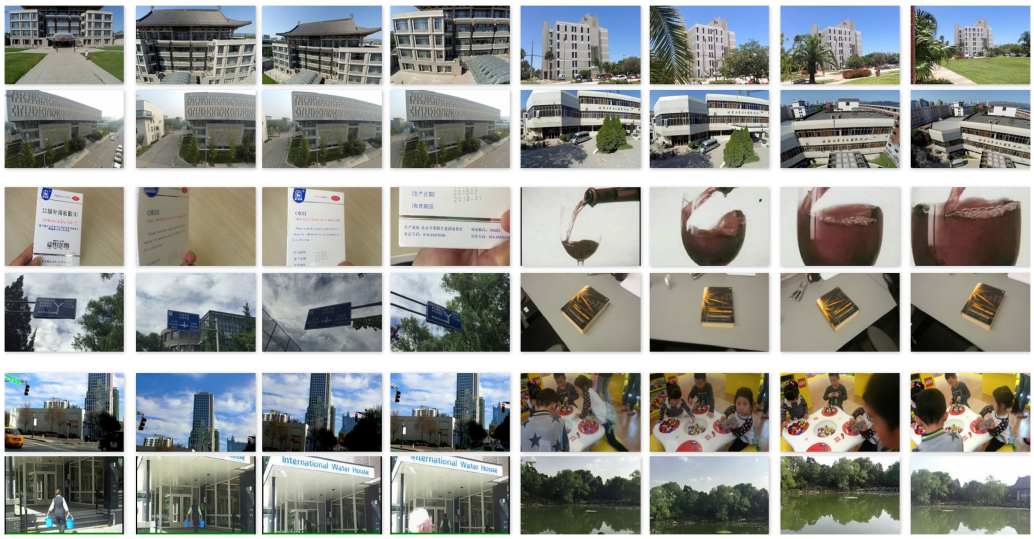

Figure 4. Examples in the MPEG-CDVA dataset.

Evaluation Framework

The MPEG-CDVA dataset includes 9974 query and 5127 reference videos, and each video takes from 1 s to 1+ min durations.[15] In Figure 4, we provide some typical examples from the MPEG-CDVA dataset. In total, 796 items of interest in those videos are depicted, which can be further divided into three categories, including large objects (e.g., buildings, landmarks), small objects [e.g., paintings, books, committee draft (CD) covers, products], and scenes (e.g., interior scenes, natural scenes, multicamera shots). Approximately 80% of query and reference videos were embedded in irrelevant content (different from those used in the queries). The start and end embedding boundaries were used for temporal localization in video matching task. The remaining 20% of query videos were applied with 7 modifications (text/logo overlay, frame rate change, interlaced/progressive conversion, transcoding, color to monochrome and contrast change, add grain, display content capture) to evaluate the effectiveness and robustness of the compact video descriptor representation technique. As such, 4693 matching pairs and 46 930 nonmatching pairs are created. In addition, for large-scale experiments 8476 videos with a total duration of more than 1000 hours are involved as distracters, which belong to UGC, broadcast archival and education.

The pairwise matching performance is evaluated in terms of the matching and localization accuracy. In particular, the matching accuracy is accessed by the receiver operating characteristic curve. The true positive rate (TPR) given false positive rate (FPR) equaling to 1% is also reported. When a matching pair is observed, the localization accuracy is further evaluated by the Jaccard Index based on the temporal location of the item of interest within the video pair. In particular, it is calculated by $\frac{[T_{\text{start}},T_{\text{end}}]\cap[T'_{\text{start}},T'_{\text{end}}]}{[T_{\text{start}},T_{\text{end}}]\cup[T'_{\text{start}},T'_{\text{end}}]}$, where $[T_{\text{start}}, T_{\text{end}}]$ denotes the ground truth and $[T'_{\text{start}}, T'_{\text{end}}]$ denotes the predicted start and end timestamps. The retrieval performance is evaluated by mean Average Precision (mAP), and moreover, precision at a given cut-off rank R for query videos (Precisian@R) is calculated. Here, R is set to be 100. As the ultimate goal is to achieve compact feature representation, the feature bitrate consumption is also measured.

Timeline and Core Experiments (CE)

The CfP of MPEG-CDVA was issued in the 111th MPEG meeting at Geneva in Feb. 2015, and subsequently responses are evaluated in Feb. 2016. According to the timeline, the CD was released in Jan. 2018 then followed by Draft of International Standard. In Oct. 2018, final draft international standard was finalized and submitted for approval and publication. There were six CE in the exploration of the MPEG-CDVA standard. The first CE investigates the temporal sampling strategy to better understand the impact of keyframes and

Table 1. Performance comparisons with the evolution of CXM models.

	Operating Point	CXM0.1	CXM0.2	CXM1.0
mAP	16 KB/s	0.660	0.721	0.721
	64 KB/s	0.673	0.727	0.727
	256 KB/s	0.680	0.730	0.730
Precisian@R	16 KB/s	0.655	0.712	0.712
	64 KB/s	0.666	0.718	0.718
	256 KB/s	0.674	0.722	0.722
TPR@FPR =0.01	16 KB/s	0.779	0.836	0.836
	64 KB/s	0.790	0.843	0.843
	256 KB/s	0.800	0.846	0.846
	16 256 KB/s	0.786	0.838	0.838
Localization accuracy	16 KB/s	0.365	0.544	0.662
	64 KB/s	0.398	0.567	0.662
	256 KB/s	0.411	0.579	0.652
	16 256 KB/s	0.382	0.542	0.652

densities in video analysis. The second CE targets at improving the matching and retrieval performance based on the segment level representation. The CE3 exploits the temporal redundancies of feature descriptors to further reduce the bitrate for feature representation. CE4 investigates the combination strategy of traditional handcrafted and deep learning based feature descriptors, and CE5 develops compact representation methods of the deep learning based feature descriptors. Finally, CE6 study the approaches for deep learning model compression to reduce the runtime and memory footprint for deep learning based feature extraction.

Performance Results

In this section, we report the performance results of the key contributions in the development of CDVA. First, the performance comparisons with the evolution of CXM models are presented. CXM 0.1 (released on MPEG-114) is the first version of CDVA experimentation model that provides the baseline performance, and subsequently, CXM0.2 (MPEG-115) and CXM1.0 (MPEG-116) have been released. To flexibly adapt to different bandwidth requirements as well as application scenarios, three operating points in terms of the feature descriptor bit rate 16 KB/s, 64 KB/s, and 256 KB/s are defined. Besides, in the matching operation, an additional cross mode 16 256 KB/s matching has also been considered. In Table 1, the performance comparisons from CXM0.1 to CXM1.0 are listed. The performance improvements from CXM0.1 to CXM0.2 are significant, and more than 5% on mAP and 5% in terms of TPR@FPR are observed, which are mainly attributed to keyframe sampling based on the color histogram. Comparing CXM0.2 with CXM1.0, the retrieval performance is identical since the changes lie in the video matching operation, which improves the localization performance based on the video shot to identify the matching interval. Such matching scheme leads to more than 10% temporal localization performance improvement.

In Table 2, the performance comparisons between CXM and deep learning based methods are provided. Compared with CXM1.0, simply using the deep learning based feature descriptors in 512 dimension without reranking techniques can bring about 5% improvements on both mAP and TPR. It can be seen that the performance of NIP descriptor extracted from a

Table 2. Performance comparisons between handcrafted and deep learning based methods.

	mAP	Precisian@R	TPR@FPR=0.01	Localization Accuracy
CXM0.1	0.66	0.655	0.779	0.365
CXM0.2	0.721	0.712	0.836	0.544
CXM1.0	0.721	0.712	0.836	0.662
NIP	0.768	0.736	0.879	0.725
NIP+SCFV	**0.826**	**0.803**	**0.886**	**0.723**
NIP (compressed model)	0.763	0.773	0.87	0.722
NIP (compressed model) + SCFV	0.822	0.798	0.878	0.722
Binarized NIP	0.71	0.673	0.86	0.713
Binarized NIP+ SCFV	0.799	0.775	0.872	0.681

compressed model only suffers a negligible loss while the model size has been reduced from 529.2M to 8.7M using pruning and scalar quantization. To meet the large-scale fast retrieval demand, the performance of binarized NIP (occupying only 512 bits) and its combination with handcrafted feature descriptors are also explored. Compared with CXM1.0, the additional 512 bits deep learning based descriptor in the combination mode significantly boosts the performance from 72.1% to 79.9%. It is worth noting that the NIP descriptor has been integrated into CXM in MPEG 119th meeting in Jul. 2017.

In Table 3, we list the runtime complexity between CXM1.0 and the deep learning based methods. In the experimental setup, for each kind of feature descriptor, the database is scanned once to generate the retrieval results. CXM1.0 adopts SCFV descriptor to obtain an initial top 500 results and then local descriptor reranking is applied. The fastest method is binarized NIP that takes 2.89 s to implement a video retrieval request in 13603 videos (about 1.2 millon keyframes), and NIP descriptor takes 9.15 s to complete this task. For handcrafted descriptor, the CXM1.0 takes 38.63 s, including both global ranking with SCFV and reranking with local descriptors. It is worth mentioning that here CDVA mainly focuses on the performance improvement in terms of the accuracy of matching and retrieval. Regarding the retrieval efficiency, some techniques such as advanced indexing methods that have not been standardized in CDVS can significantly improve the retrieval speed, and they have not been integrated into CDVA for investigation either.

Table 3. Runtime complexity comparisons between CXM1.0 and the deep learning based methods.

	Retrieve query(s/q.)	Matching pair (s/p.)	Non Matching piar(s/p.)
CXM1.0	38.63	0.37	0.21
NIP	9.15	0.3	0.26
NIP+SCFV	39.45	0.38	0.29
Binarized NIP	2.89	0.29	0.25
Binarized NIP +SCFV	33.24	0.37	0.28

Computation platform: 2 processors, 2x12 cores, Xeon E5-2692V2@2.2 GHZ, 64 GB RAM.

CONCLUSION AND OUTLOOK

The current development of CDVA treats the CDVS as the groundwork, as they serve the same purpose of using compact feature descriptors for visual search and analysis. The main difference lies in that CDVS is mainly focusing on still images, while CDVA makes an extension to video sequences. Moreover, the backward compatibility of CDVA supports the feature decoding of the keyframe with the existing CDVS infrastructure, such that every standard compatible CDVS decoder can reproduce the features of independently coded frames in the CDVA bitstream.

This can greatly facilitate the cross-modality search applications, such as using images as queries to search videos or using videos as queries to search corresponding images.

The remarkable technological progress in video feature representation has provided a further boost for the standardization of compact video descriptors. The key frame representation and interfeature prediction provide two granularity levels in video feature representation. The deep learning feature descriptors have also been intensively investigated, including the feature extraction, model compression, compact feature representation, and the combination of deep learned based features with traditional handcrafted features. The optimization of the video matching and retrieval pipelines has also been proved to bring superior performance in video analysis.

Nevertheless, the standardization of CDVA is also facing many challenges and more improvements are expected. In addition to video matching and retrieval, more video analysis tasks (such as action recognition, abnormal detection, video tracking) need to be investigated. This requires more advanced video representation techniques to extract the motion information as well as sophisticated deep learning models with high generalization ability for feature extraction. Moreover, although the deep learning method has achieved significant performance improvement, more deep feature compression and hashing work are necessary to achieve compact representation. In addition, the fusion strategy of deep learning feature and traditional handcrafted feature pose new challenges to the standardization of CDVA and open up new space for future exploration. Finally, normatively describing the CNN based descriptors is indeed crucial in standardization. Regarding the standardization of the deep neural network, the convolution layer, pooling layer, and the region sampling parameters involved in feature generation stage should be specified to enable the interoperability. In the future, more experiments are also expected to be conducted regarding the aforementioned descriptions standardization, from the perspectives of the analysis accuracy and complexity.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China under Grant 2016YFB1001501, in part by the National Natural Science Foundation of China under Grant 61661146005 and Grant U1611461, and in part by the National Research Foundation, Prime Minister's Office, Singapore, under the NRF-NSFC Grant NRF2016NRF-NSFC001-098. L.-Y. Duan is the corresponding author.

REFERENCES

1. "Information technology on multimedia content description in- terface part 13: Compact descriptors for visual search," ISO/IEC 15938-13:2015, 2015.
2. L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans.Image Process.*, vol. 26, no. 20, pp. 179–194, Jan. 2016.
3. "Call for proposals for compact descriptors for video analysis (CDVA)-search and retrieval," ISO/IEC JTC1/SC29/WG11/N15339, Warsaw, Poland, Jun. 2015.
4. "Compact descriptors for video analysis: Objectives, applications and use cases," ISO/IEC JTC1/SC29/WG11/N14507, Valencia, Spain, Mar. 2014.
5. M. Balestri, M. Bober, and W. Bailer, "Cdva experimentation model (cxm) 0.1," ISO/IEC JTC1/SC29/WG11/N16064, San Diego, CA, USA, Feb., 2016.
6. M. Balestri, G. Francini, S. Lepsoy, M. Bober, and S. Husain, "Bridget report on cdva core experiment 1 (ce1)," ISO/IEC JTC1/SC29/WG11/M38664, Geneva, Switzerland, May 2016.
7. M. Balestri, M. Bober, and W. Bailer, "Cdva experimentation model (cxm) 0.2," ISO/IEC JTC1/SC29/WG11/N16274, Geneva, Switzerland, May 2016.
8. W. Bailer, S. Wechtitsch, and M. Thaler, "Compressing visual descriptors of image sequences," in *Proc. Int. Conf. Multimedia Model*, 2017, pp. 124–135.
9. Z. Huang, L. Wei, S. Wang, L. Duan, and J. Chen, "Pku response to core experiment 1," ISO/IEC JTC1/SC29/WG11/M38625, Geneva, Switzerland, May 2016.
10. V. Lempitsky, A. Babenko, "Aggregating local deep fea- tures for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1269–1277.
11. G. Tolias, R. Sicre, and H. e´ Je´gou, "Particular object retrieval with integral max-pooling of CNN activations," arXiv:1511.05879, 2015.
12. Y. Lou *et al.*, "Compact deep invariant descriptors for video retrieval," in *Proc. Data Compression Conf.*, 2017, pp. 420–429.

13. Y. Bai *et al.*, "PKU's Response to CDVA CE4: NIP Network Compression," ISO/IEC JTC1/SC29/WG11/m39853, Geneva, Switzerland, Jan. 2017.
14. M. Balestri, G. Francini, and S. Lepsoy, "Improved temporal localization for CDVA," ISO/IEC JTC1/SC29/WG11/M39433, Chengdu, China, Oct. 2016.
15. "Evaluation framework for compact descriptors for video analysis – Search and retrieval," ISO/IEC JTC1/SC29/WG11 N15338, 2015.

Ling-Yu Duan is currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China. He was the Associate Director with the Rapid-Rich Object Search Laboratory, a joint lab between Nanyang Technological University, Singapore, and PKU since 2012. Contact him at lingyu@pku.edu.cn.

Vijay Chandrasekhar received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2013. His research interests include mobile audio and visual search, large-scale image and video retrieval, machine learning, and data compression. His Ph.D. work on feature compression led to the MPEG-CDVS (Compact Descriptors for Visual Search) standard, which he actively contributed from 2010 to 2013. Contact him at vijay@i2r.a-star.edu.sg.

Shiqi Wang received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree in computer application technology from the Peking University, Beijing, China. He is currently an Assistant Professor in the Department of Computer Science, City University of Hong Kong, Hong Kong. Contact him at shiqwang@cityu.edu.hk.

Yihang Lou received the B.S. degree in software engineering from Dalian University of Technology, Liaoning, China, in 2015, and is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current interests include large-scale video retrieval and object detection. Contact him at yihanglou@pku.edu.cn.

Jie Lin is currently a Research Scientist with the Institute of Infocomm Research, A*STAR, Singapore. He was previously a visiting student with Nanyang Technological University, Singapore, and the Institute of Digital Media, Peking University, Beijing, China, from 2011 to 2014. His work on image feature coding has been recognized as core contribution to the MPEG-7 compact descriptors for visual search standard. Contact him at lin-j@i2r.a-star.edu.sg.

Yan Bai received the B.S. degree in software engineering from Dalian University of Technology, Liaoning, China, in 2015, and is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. Her research interests include large-scale video retrieval and fine-grained visual recognition. Contact her at yanbai@pku.edu.cn.

Tiejun Huang is a Professor and the Chair of the Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. His research areas include video coding, image understanding, and neuromorphic computing. He is a Member of the Board of the Chinese Institute of Electronics and the Advisory Board of IEEE Computing Now. Contact him at tjhuang@pku.edu.cn.

Alex Chichung Kot (F'06) is currently a Professor with the College of Engineering and the Director of the Rapid-Rich Object Search Laboratory. He has authored or coauthored extensively in the areas of signal processing for com- munication, biometrics, data-hiding, image forensics, and information security. He is a Member of the IEEE Fellow Evaluation Committee and a Fellow of Academy of Engineering, Singapore. Contact him at eackot@ntu.edu.sg.

Wen Gao (F'09) is currently a Professor in computer science with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. He has chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. Contact him at wgao@pku.edu.cn.