Feature Article

AI-Oriented Large-Scale Video Management for Smart City: Technologies, Standards, and Beyond

Lingyu Duan
Peking University, Peng Cheng Laboratory

Yihang Lou
Peking University

Shiqi Wang
City University of Hong Kong

Wen Gao
Peking University, Peng Cheng Laboratory

Yong Rui
Lenovo Research

***Abstract*—Deep learning has achieved substantial success in intelligent video analysis. To practically facilitate deep neural network models in the large-scale video analysis, there are still unprecedented challenges. Deep feature coding, instead of video coding, provides a practical solution for handling the large-scale video surveillance data. To enable interoperability in the context of deep feature coding, standardization is urgent and important. This paper envisions the future deep feature coding standard for the AI-oriented large-scale video management and discusses existing techniques, standards, and possible solutions for these open problems.**

Recently, a considerable number of deep learning algorithms have been proposed, which exhibit substantial performance improvement in various computer vision tasks. Compared with traditional handcrafted features, deep learning algorithms aim to learn representative features from the vast amounts of training data. After AlexNet[1] won the ImageNet competition, there are tremendous research activities focusing on designing more powerful and deeper networks. Follow ups like GoogleNet,[2] ResNet[3] have greatly improved the discrimination capability of features to a higher level, which also boosted the performance of many visual analysis tasks. Generally speaking, these technologies have naturally made

Digital Object Identifier 10.1109/MMUL.2018.2873564
Date of publication 25 October 2018; date of current version 12 June 2019.

1070-986X © 2018 IEEE

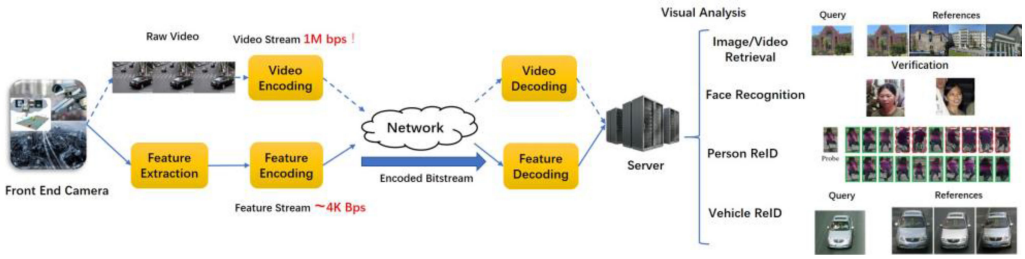

Figure 1. Infrastructure of the large-scale video management with feature transmission for smart city applications.

substantial impact on public security, such as face recognition,[4] person,[5] and vehicle reidentification (ReID)[6] in surveillance videos.

Recent years have witnessed dramatically increased demand for the smart city construction, where the concerning safety issues have received sufficient interest. In particular, there is a vast and increasing proliferation of surveillance videos acquired and transmitted over both wireline and wireless networks. Due to the real-time recording of the physical world, surveillance video is very valuable, and there is considerable concern regarding how to efficiently manage such surveillance video big data. In view of the explosion of the surveillance systems deployed in urban areas and millions of objects/events captured every day, there are a unique set of challenges regarding efficient analysis and search. In particular, video compression and transmission constitute the basic infrastructure to support these applications. Though the state-of-the-art video coding standards such as H.265/HEVC have dramatically improved the coding performance, it is still questionable that whether such big video data can be efficiently handled by visual signal level compression. Fortunately, an alternative strategy "analyze then compress" provides a solution, which transmits the compact features extracted and compressed at the edge end to the server side. Such a paradigm can sufficiently satisfy various intelligent video analysis tasks, by using significantly less data than the compressed video itself. In Figure 1, the infrastructure of the smart city with large-scale video management based on feature extraction and transmission is illustrated. In particular, to meet the demand for large-scale video analysis in smart city applications, the feature stream instead of video signal stream can be transmitted. As such, the intelligent front-end devices extract features locally and then convey the encoded feature stream to the server for analysis purpose.

While the field of artificial intelligence is still quickly evolving and efficient and novel deep learning algorithms will continue to emerge in the coming years, it is also interesting to discuss how we could enable the interoperability of the compressed deep learning features in real-world applications. In contrast with video coding, which directly compresses the visual signals into the bitstream, feature coding involves both feature extraction and compression process. In particular, feature extraction serves as the raw features producer to generate the source for compression and is responsible for the answer of what to compression. Feature compression accounts for the conversion of raw deep features into compact representation bitstream. The purpose of this paper is to provide an overview of the existing deep learning techniques in video surveillance and envision the future deep learning feature coding standards. We will start by a brief review of the current status of deep learning in video surveillance, followed by discussions on the compact feature standard in MPEG. Then, the open problems of deep feature coding standardization will be discussed, where we can perceive both great promises and challenges.

STANDARDIZATION OF A COMPACT HANDCRAFTED FEATURE DESCRIPTOR

The compact descriptors standard from ISO/IEC moving pictures experts group has succeeded in enabling the interoperability for efficient and effective image/video retrieval by standardizing the bitstream syntax of compact feature descriptors. Over the course of the standardization process, remarkable improvements

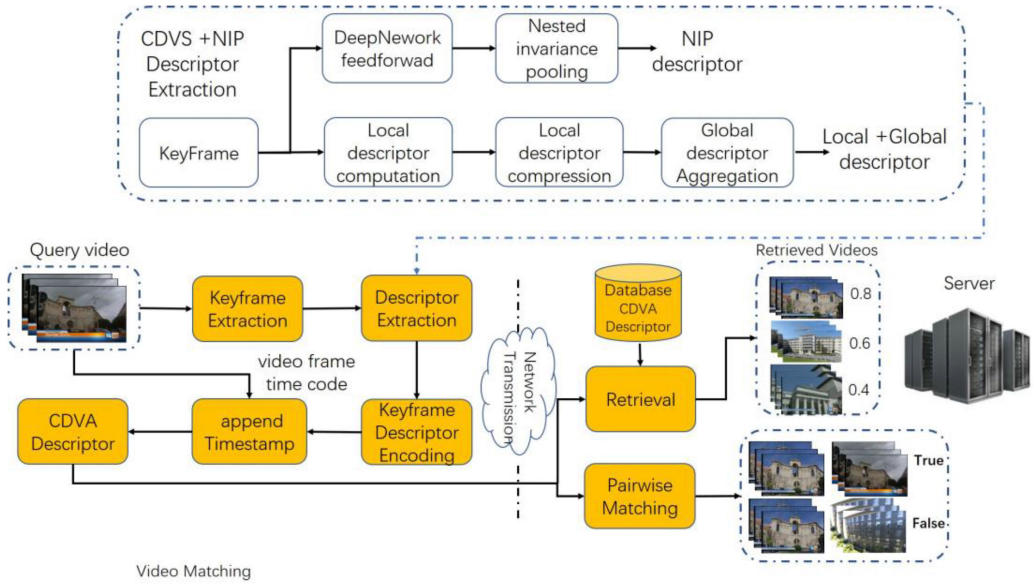

Figure 2. Illustration of the MPEG CDVA evaluation framework.

were achieved in reducing the size of feature data and in reducing the computation and memory footprint in the feature extraction process.

Compact Descriptors for Visual Search

In view of the importance of the transmission of feature descriptors, MPEG has finalized the standardization of compact descriptors for visual search (CDVS)[7] (ISO/IEC15938-13) and published the standard in Sep. 2015. In CDVS, handcrafted local and global descriptors are leveraged to represent the visual characteristics of images. The normative blocks of CDVS involve the extraction of local and global descriptors. More specifically, the local descriptors consist of SIFT descriptors, which are efficiently compressed by a low-complexity transform coding. The raw local descriptors are further selected and aggregated to generate a scalable compressed Fisher vector, with competitive matching accuracy and low memory footprint. In view of the fluctuation of available bandwidth in the mobile environment, CDVS supports interoperability between different size bitstream by setting six operating points from 512 B to 16 KB. More technical details about CDVS are presented by Duan *et al.*[7]

Compact Descriptors for Video Analysis

The emerging requirements of video analysis facilitate the standardization of large-scale video analysis. The MPEG has moved forward to standardize compact descriptors for video analysis (CDVA).[8] Video consists of sequentially correlated frames, such that extracting the feature from each frame leads to high redundancy in feature representation and unnecessary computational costs. The ongoing CDVA standard adopted multikeyframe-based image retrieval, which converts the problem of video retrieval into an image retrieval task. In particular, the local and global descriptors of standardized CDVS descriptors are extracted on the sampled keyframes of a given query video, which are further packed together to constitute the CDVA descriptors. Moreover, the deep learning features[9] also adopted to further boost the analysis performance. Figure 2 presents the framework of ongoing CDVA with handcrafted features and deep features, including the video structure and the normative components of feature extraction. It is also worth mentioning that the NIP descriptor has been adopted into the CDVA standard (ISO/IEC15938-15).

DEEP LEARNING IN VIDEO SURVEILLANCE

With the exponential growth of the video surveillance data, video content analysis has been a long-standing research topic in computer vision community. Currently, there are four urgent visual analysis tasks in the surveillance scenario,

Table 1. Core techniques involved in the visual system of a smart city.

Core techniques	Description
Feature generation	Extract discriminative feature representation
Feature generalization	Generalize the deep feature, to different tasks, e.g., from person ReID to vehicle ReID.
Feature redundancy removal	Remove the redundancies of the features in spatial or temporal domain
Rate-distortion optimization	Investigate the distortion of features and optimize feature compression with rate-distortion optimization.
Feature binarization	Binarize features to enable fast feature transmission and analysis.
Network compression	Efficiently represent the network and lower the disk and memory cost.

i.e., image/video retrieval, person re-identification, face recognition, and vehicle retrieval. These tasks play important roles in building the safety city and ensuring the public security. Associated with these tasks, the core techniques in establishing the visual system of the smart city are shown in Table 1. Multiple academic disciplines, including visual signal processing, computer vision, compression, as well as hardware architectures, are involved in the further construction of the system. It is envisioned that with the standardization of deep learning features and the advancements of these technologies, the system that sees intelligently, efficiently, and greenly will eventually come true.

Image/Video Retrieval

Image/video retrieval refers to searching for the images/videos representing the same objects or scenes as the one depicted in the query, which may present under different scales, illuminations, rotations, or even occlusions. In the last decade, the image/video retrieval has benefited a lot from handcrafted SIFT descriptors due to its robustness to the image transformations. However, after the AlexNet[1] won ILSVRC12 by a significant margin, the CNN-based image feature representation has become mainstream techniques when handing complex and semantic vision analysis. Regarding both image and video retrieval, competitive and even better retrieval performance[9–11] has been reported on several benchmarks.

CNN-based retrieval methods can be categorized into two types: pretrained and fine-tuned CNN models. The commonly used pretrained CNN models are trained on ImageNet dataset consisting of 1.2 million images of 1000 classes, such that the features can be regarded as generic. The descriptors can be extracted from fully connected (FC) layers or intermediate layers. The FC descriptors have a global receptive field and the intermediate local descriptors have a smaller receptive field and location information encoded in 2-D feature maps. To obtain the global representation, encodings like VLAD and FV are usually adopted. In addition, the direct pooling can also generate discriminative features. For example, Liu *et al.*[11] employed max pooling on selected regions in intermediate feature maps and subsequently performed sum pooling. Though impressive results can be achieved by the pretrained model, there is a trend to fine-tune a CNN model on a task-oriented dataset for specific retrieval. The classification- and verification-based networks are two typical types. The former is trained to classify predefined categories, and the latter adopts siamese network[12] with contrastive loss or triplet loss. On several retrieval benchmarks such as Holidays, Oxford5K, and Paris 6K, the fine-tuned models have achieved the state-of-the-art performance. The evaluation criteria in image/video retrieval are the widely used mean average precision (mAP). We investigate the performance of relevant methods on an MPEG-CDVA dataset. The MPEG CDVA dataset includes 9974 query and 5127 reference videos, which contain large objects (e.g., buildings and landmarks), small objects (e.g., books and products), and scenes (e.g., interior or natural scenes). For retrieval experiments, 8476 videos with more than 1000 h (about 1.2 million keyframes) in terms of user-generated content, broadcast are used as distractors.

Person Reidentification

Person ReID has attracted more and more research focus due to its application significance in video surveillance. It aims to search whether a given person is present in other cameras. The

widespread camera deployment in public places and the increased safety requirements make the existing manual labor spotting scheme powerless when facing the real-time generated massive video data. A practical person ReID system involves person detection, tracking, and retrieval. In particular, person retrieval is the main research focus among the related works. The challenge of this task is how to accurately match two images of the same person under variant scenes, viewpoints, scales, and lighting conditions.

Deep ReID systems mostly employ two types of CNN models, i.e., siamese and classification models. The difference between these two models lies in the input form and definition of the loss function. The siamese models[13] leverage image pairs or triplets as input, and then let them forward propagate to get feature vector in embedding space. The distance of images with the same person is constrained on the feature vector by a minimum margin using contrastive loss or triplet loss. Liu *et al.*[11] proposed a multiscale network consisting of deep and shallow subnetworks, which is able to learn discriminative person representation at various spatial scales. By contrast, the classification models proposed by Zhao *et al.*[14] treat each person identity as a class, and the classification-based loss functions are usually employed, such as softmax loss. These models pay more attention to the feature representation from the perspective of local and global combination, part-attention model, human body's skeleton model, etc. Intuitively, the combinations of siamese and classification model have also received a lot of attention. The performance of person ReID algorithms is mostly evaluated by mAP and precision @R. We explore the methods reported on four representative datasets in person ReID, i.e., CUHK1, VIPeR, PRID, and GRID dataset. These four datasets cover different scales application and scenarios, and all of them involve multicamera ReID.

Face Recognition

Due to the nonintrusive recognition manner (intrusive like fingerprint and retina recognition), face recognition has great application potential in surveillance security. Over the last decades, a number of works in face recognition[4,15] have emerged, which greatly boosted the accuracy on the popular benchmark such as labeled face in the wild (LFW) to an unprecedented level. In real-world applications, the captured face images may not be as high quality as that in LFW dataset, creating many challenging problems originated from arbitrary poses, low-quality resolutions, occlusions, and small scales.

Typically, face recognition includes face identification and face verification. The former classifies a given face as a specific identity, and the latter verifies whether a given face pair belongs to the same identity. Regarding the experimental setup, there are closed-set and open-set settings. Under the closed set, the testing identities are contained in the training set. By contrast, in the open set, the testing identities do not appear in the training set. Therefore, the real-world recognition can be regarded as face verification in the open set. Essentially, this task is defined as a metric learning problem. The expected feature representation should be able to meet the demand for small intraclass distance and large interclass distance. Deep models are capable of building the above-mentioned criterion by setting appropriate loss functions. Finally, the matching score of face recognition is computed by the cosine distance of two features. The nearest neighbor classifier and thresholding are used for face identification and verification, respectively. In this paper, we discuss the methods reported on the LFW and YouTube faces (YTF) datasets. LFW contains more than 13 000 images of 5749 persons from the web. YTF dataset contains 3425 videos of 1595 different people.

Vehicle Reidentification

In many vehicle-relevant tasks, vehicle ReID is the most crucial technique in city security. The license plate is usually the straightforward choice to identify a vehicle. However, in real applications, most surveillance cameras are not equipped with recognition capability. Furthermore, the license plates of vehicles in many cases are occluded or faked. Thus, the visual appearance based techniques present great application prospect. Compared with the classic person ReID problem, vehicle ReID is more challenging since it faces the enormous interclass similarity and intraclass variances presented by

Table 2. Performance comparisons of methods reported on CDVA benchmarks where the landmarks, scene, and objects are the subdatasets (Refer to papers by Balestri *et al.*[17–45] for references).

Methods	Dims	Year	Land mark	Scene	Objects	All
CXM0.2	1024	2016	0.598	0.594	0.917	0.721
MAC	512	2015	0.619	0.762	0.718	0.67
SPoC	256	2015	0.691	0.84	0.703	0.709
CroW	512	2015	0.639	0.784	0.72	0.683
R-MAC	512	2015	0.746	0.873	0.782	0.771
HNIP VGG	512	2016	0.748	0.901	0.85	0.801
HNIP Alex	768	2016	–	–	–	0.772
HNIP Res	2048	2016	–	–	–	0.817

massive vehicles of the same model types and the shooting situation variations across multiple cameras. For example, the subtle differences between similar vehicles are even challengeable for human beings. Fortunately, some special marks such as tissue box, pendant, annual inspection marks, etc., provide characteristics clues for efficient discrimination.

The deep metric learning has been widely adopted for vehicle ReID tasks. The objective of the deep network is to learn a deep embedding where the samples of the same vehicle ID are constrained in a local space, such that the samples of different vehicle ID are farther away than ones of the same vehicle ID. Such feature distribution is pretty desirable for nearest neighbor retrieval. As such, the retrieval method is the main solution for ReID. As mentioned above, the granularity of vehicle ReID is finer than person ReID. Consequently, the interclass feature distribution requires more structured prior knowledge to represent such subtle differences. This motivated some recent attempts, which incorporate intraclass variance into the feature representation, such as group-sensitive triplet embedding by Bai *et al.*[16] For vehicle retrieval performance evaluations, mAP and mean precision @K are widely used. Moreover, for ReID evaluation, the widely used metric is cumulative match curve. We explore the methods on two recently published vehicle ReID dataset VehicleID and VERI-776 datasets. The former contains 221 763 with 26 267 vehicle images. The latter one consists of vehicle images about 50 000 images of 776 vehicles. These two datasets are both collected from real-world traffic scenario, in which each vehicle is captured in different viewpoints, illuminations, and resolutions.

STANDARDIZATION OF A DEEP FEATURE DESCRIPTOR

In the context of video big data, to further ensure interoperability in deep learning based video analysis, a standard that focuses primarily on defining the syntax of compressed deep feature descriptors is essential. This section clarifies the issues to be solved in the standardization process and how they might be pragmatically approached. We believe that such AI-oriented standard could represent a sea change in the future smart city applications.

Compact Deep Feature for Video Analysis

As introduced in "Deep Learning in Video Surveillance," the features extracted by deep neural networks are gradually replacing the handcrafted features in many visual intelligence analyses. Due to millions of parameters lying in the deep network, as well as a series of nonlinear mappings, the deep network can present high discrimination capability with pretty lower memory costs compared with the handcrafted features. Moreover, when massive training data are available, the involvement of an end-to-end learning scheme would further sharpen the feature discrimination ability. Here, we investigate the performance and feature compactness of the

Table 3. Performance comparisons of vehicle re-identification methods on VehicleID and VeRI benchmarks (Refer to papers by Balestri *et al.*[17–45] for references).

Methods	Dims	Year	VEHICLEID	VERI-776
Triplet	400	2015	0.373	–
Softmax Loss	1024	2015	0.580	0.343
Triplet+Softmax	1024	2014	0.650	0.558
BOW-CN	100	2015	–	0.122
CCL VGGM	1024	2016	0.386	–
Mixed Diff+CCL	1024	2016	0.455	–
HDC+Contrastive	384	2017	0.575	–
GSTE	1024	2017	0.724	0.594

recent remarkable works in four typical analysis tasks in city surveillance.

Although different network structures are employed in different analysis tasks, we find that the features can be uniformly represented without significantly sacrificing the analysis accuracy. Table 2 lists the video retrieval performance of the deep learning features with off-the-shelf CNN model reported in CDVA benchmarks. From the perspective of the performance and feature compactness, the deep learning feature shows the competitive performance. With the advance of network structure and training scheme, the performance of deep features has also been dramatically improved, being state-of-the-art on several image retrieval benchmarks, such as Holiday, Oxford5K, and Paris6K. In image retrieval applications, the scale, translation, and rotation changes greatly affect the feature representation. The recent deep based method such as R-MAC and HNIP mainly focus on the invariant pooling to generate deep invariant features. The person/vehicle ReID tasks also benefit a lot from the success of deep networks, as shown in Tables 3 and 4. Generating discriminative feature distribution is crucial to the performance of vehicle and person ReID. In these two tasks, the proposed methods such as mixed Diff, HDC, and GSTE all

Table 4. Performance summarization of some representative Person ReID methods on the benchmarks (Refer to papers by Balestri *et al.*[17–45] for references).

Methods	Dims	Year	VIPeR	CUHK1	PRID	GRID
Cov-of-Cov	16828	2016	33.9	40.9	47	16.6
LOMO	26960	2015	40	–	15.3	16.6
GOLD	1169	2015	27.1	35.3	40.5	10.9
2AvgP	952	2015	28.8	36.1	44.7	12.9
GOG-RGB	7567	2015	42.3	55.8	63.6	22.8
NFST	5138	2016	51.2	69	–	–
SCSP	120	2016	53.5	24.2	–	–
SSDAL	105	2016	43.5	–	20.1	19.1
TMA	100	2016	39.9	–	54.2	–
P2S	800	2017	–	77.3	–	–
Spindle	256	2017	53.8	77.9	67	–

target at maximizing interclass distances and meanwhile minimizing intraclass distances in feature spaces. Specifically, in the vehicle ReID task, different vehicles with the same model types are considered as different class IDs. The only cues can be utilized are the characteristics signs on vehicles, which require the algorithm to be enough sensitive to these subtle differences. As for face recognition, the recent method such as sphereface tends to generate discriminative feature distribution by applying improved softmax loss. In particular, we observe there is a trend that the recent methods employ features with much lower dimensions to produce the representation of an object, with the help of more powerful networks and well-defined loss functions. Similar trends can also be found in the recent efforts for face recognition, as shown in Table 5. In addition, the performance improvements originating from introducing object proposals in image retrieval have been witnessed in several image retrieval benchmarks such as Oxford and Paris. The performance gains originate from the better localization ability provided by region proposal, as shown in Figure 3(c) and (d). There is a tradeoff between numbers of proposals and complexity. As for these multiple-target retrieval methods, generating efficient and accurate proposal boxes can be beneficial to retrieval performance. However, the obvious demerit of region-based methods is the relatively high feature dimension due to the representation requirements of multiple targets in query or reference images. As such, the region proposal mechanism would be practically applicable with the representations of the compact feature. In Figure 3(a) and (b), the performance variations against the dimension are illustrated. All these experimental results demonstrate that these tasks can be successfully achieved with identical or similar feature dimension. For example, when the feature size reaches 512 dimensions, in most of the face recognition cases, competitive results can be obtained. Obviously, a similar phenomenon can also be found when the dimension reaches 512 in CDVA, 512 in person ReID, and 1024 in vehicle ReID. In a word, a converging point can be feasibly attained from the perspectives of feature dimensions, proposal numbers, and network structure. Another observation is that the performance variation along with the augment of proposals can also arrive at a saturation eventually, as shown in Figure 3(c) and (d). Such observations provide useful evidence for the further standardization of deep learning features, as discussed in "Toward Standardization of Deep

Table 5. Summary of efficiency and accuracy comparisons between recent remarkable works on face recognition (Refer to papers by Balestri *et al.*[17–45] for references).

Methods	Year	Dims	LWF	YTF
DeepFace	2014	4096	97.35	91.4
Learning face	2014	10575	97.73	92.2
MDML-DCPs	2015	1024	98.95	97.3
FaceNet	2015	128	99.63	95.12
Deep embedding	2015	128	99.13	–
Multimodal face	2015	9000	98.43	–
Center loss	2016	512	99.28	94.9
Large-margin softmax	2016	512	98.71	–
SphereFace	2017	512	99.42	95
Neural aggregation	2017	128	–	95.72

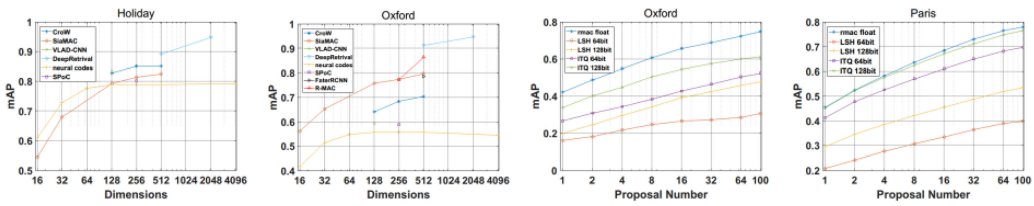

Figure 3. Performance variations with the increase of dimensions and object proposals for recent works on image retrieval benchmarks.

Features." In the future, it is also expected that more compact and discriminative feature representations will emerge due to the advance of the network architectures and optimization strategies.

Toward Standardization of Deep Features

It is apparent that the compact deep feature possesses many favorable properties for the applications of the smart city. However, the explosion of the deep learning models is also creating many challenging research problems. In particular, it is worth noting that the feature coding differs with traditional video coding in that an end-to-end feature coding pipeline involves both feature extraction and compression. In other words, for video coding, the source visual signals are established and available, i.e., the pixel values. By contrast, in feature coding, different deep learning models would create dramatically different features for the subsequent compression process. Therefore, a complete and exhaustive standard that can fully ensure the interoperability typically specifies the standardization of both feature extraction and compression. As such, any bitstreams that conform to such standard can be meaningfully compared.

Such standardization requires the deterministic deep network model and parameters. Nevertheless, the recent research achievements of deep learning emerge in endlessly, and moreover, there is a lack of the generic deep model that can be applied to a broad of tasks in video surveillance. Therefore, the standardization of the deep learning model is not ready for prime time.

Here, we propose the concept of semi-interoperability for feature coding, which only standardizes the feature compression. In other words, only the pipeline from raw features to the compressed bitstreams is taken into consideration, and the final syntax that specifies the compact deep features is standardized. The raw feature extraction process is left open for future exploration. Such strategy is based on the key observation that the raw features for these tasks can be uniformly represented. As such, the increasing demand for the interoperability in the smart city and the explosion of deep learning techniques can be well balanced.

The semi-interoperability based standardization strategy is dual to the video coding standard where only the decoder is standardized. The decoder conforming to the standard can only correctly recover the features but does not account for the explanation of the features as the deep learning model is not specified. Therefore, such strategy only ensures that any deep learning feature bitstreams from the same deep learning model conforming to such standard can be meaningfully matched after decoding. In other words, it does not fully support the interoperability and bitstreams conforming to such standard may convey different information. On the other hand, the advantage lies in that in the future any effective deep learning models can seamlessly collaborate with this standard, such that the standard can be kept with long-lasting vitality. Moreover, though there are multiple tasks in video surveillance and each task corresponds to the specific deep learning model, as long as the final generalized bitstreams from these models conform to the standard, they can be successfully decoded by a unified decoder. Here, the traditional feature compression and standardized feature compression frameworks are shown in Figures 4 and 5. It is observed that the bitstreams from different ends can be uniformly represented and transmitted, such that a unified decoder can be used to decode such bitstreams to enable the semi-interoperability.

Regarding feature compression, the high redundancy of deep learning features in video sequences needs to be removed. Specifically, due to the similar content in continues video frames in a shot, the deep learning features also contain representation overlap. Therefore, the residuals change between deep features in a shot can be encoded to reduce redundancy. In particular, many video coding technologies can be analogously transferred to feature codings, such as interprediction, intraprediction, and rate distortion optimization. In addition, since the basic role of video surveillance is to analyze and explain the object behaviors, and in many occasions within a video frame there are multiple objects, it is natural to extend the frame level feature extraction and compression to the object level based on object proposals. For example, real-time object detectors such as

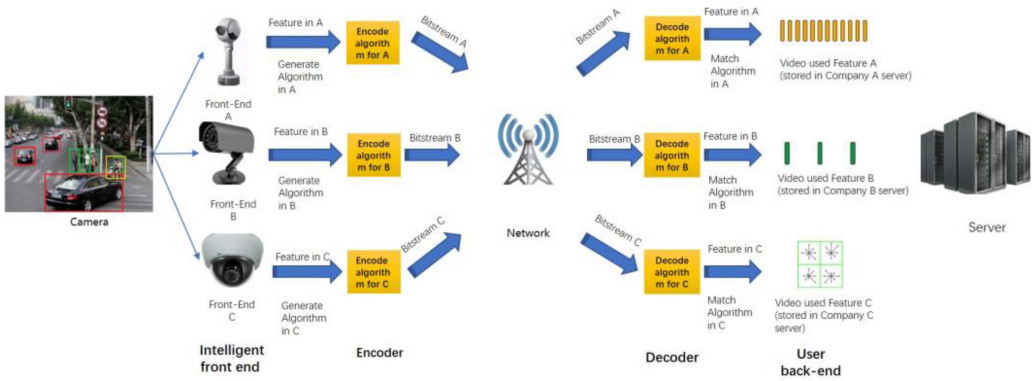

Figure 4. Illustration of traditional feature transmission framework. Different feature bitstreams originated from different front-end cameras have their own organization and syntax, such that the corresponding encoding and decoding algorithms should be performed at local and server ends, respectively.

YOLO[46] can be adopted to localize the target objects such as persons, vehicles, heads, or other objects of interest, and then the regions of interest will be fed into the corresponding networks designed for specific tasks to obtain the feature representation. This also requires the nonlocal intraprediction to remove the redundancy from different objects within a frame. As such, how these redundancies can be removed and how the final bitstream is composed of should be further investigated in the standardization exploration.

It is also anticipated that in the future, the deep learning models are developed to maturation and generic, as well as dynamic feature representations can be learned from surveillance videos. At that stage, there may emerge a unified deep learning model that can be standardized to achieve the full interoperability. Generally speaking, such deep learning model can not only deal with the various video surveillance tasks but also enjoy the properties such as lightweight and friendly for implementation. Overall, the message we are trying to send here is not that the

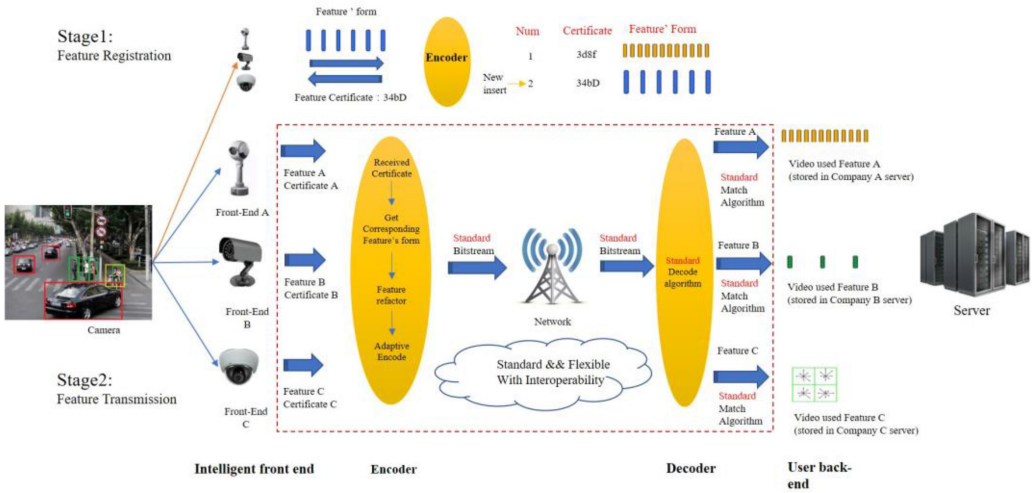

Figure 5. Illustration of the interoperability-enabled feature transmission framework. First, new features are required to register its feature organization and syntax in the encoder and obtain the corresponding certificate. Second, features generated at the front end will be reorganized according to its certificate registered in the encoder, and then encoded in the standard bitstream syntax. Therefore, the server end can leverage the standard algorithm to decode the received features.

standardization of the deep model for feature extraction is abandoned. Rather, we hope to make the point that at the current stage, there are flexible and practical alternative solutions for the standardization that can be deployed.

OUTLOOK

We have discussed the practical issues and envisioned the future standardization of deep learning features in the context of large-scale video management in the smart city. Rather than exhaustively establishing the whole feature representation process including both extraction and compression, we have emphasized on the great potentials of standardizing the bitstream syntax of the compressed features. Such strategy is significantly different from the previous MPEG-7 visual standards such as CDVS and CDVA, and the deep learning models are not required to be specified to conform to the standard, which further enhances the flexibilities in the proliferation of deep learning technologies. In the future, it is expected that such AI-oriented feature coding standard plays important roles in the establishment of the visual system of the city brain and impact the new development of future AI technologies.

ACKNOWLEDGMENT

This work was supported the National Basic Research Program of China under grant 2015CB351806, in part by the National Natural Science Foundation of China under Grant 61661146005, Grant U1611461, and in part by Hong Kong RGC Early Career Scheme under Grant 9048122 (CityU 21211018).

APPENDIX A

A.1 Methods reported on CDVA benchmarks

- CXM0.2[17]
- MAC[18]
- SPoC[19]
- CroW[19]
- R-MAC[18]
- HNIP VGG[20]
- HNIP Alex[20]
- HNIP Res[20]

A.2 Methods on VehicleID and VeRI benchmarks

- Triplet[21]
- Triplet+Softmax[21]
- CCL VGGM[22]
- Mixed Diff+CCL[22]
- HDC+Contrastive[23]
- GSTE[24]

A.3 Methods on person ReID benchmarks

- Cov-of-Cov[25]
- LOMO[26]
- GOLD[27]
- 2AvgP[28]
- GOG-RGB[29]
- NFST[30]
- SCSP[31]
- SSDAL[32]
- TMA[33]
- P2S[34]
- Spindle[35]

A.4 Methods on face recognition benchmarks

- DeepFace[36]
- Learning Face[37]
- MDML-DCPs[38]
- FaceNet[39]
- Deep embedding[40]
- Multimodal face[41]
- Center Loss[42]
- Large-margin softmax[43]
- SphereFace[44]
- Neural Aggregation[45]

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
2. C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 1–9.
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 770–778.
4. W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2017, pp. 6738–6746.
5. T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1249–1258.

6. H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 2167–2175.
7. L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 26, no. 20, pp. 179–194, Jan. 2016.
8. "Call for proposals for compact descriptors for video analysis(CDVA)-search and retrieval," ISO/IEC JTC1/SC29/WG11/N15339, Warsaw, Jun. 2015.
9. J. Lin *et al.*, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.
10. V. Lempitsky and A. Babenko, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1269–1277.
11. J. Liu *et al.*, "Multi-scale triplet CNN for person reidentification," in *Proc. ACM Multimedia Conf.*, 2016, pp. 192–196.
12. A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 241–257.
13. D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person reidentification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1268–1277.
14. H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2017, pp. 907–915.
15. Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 499–515.
16. Y. Bai, F. Gao, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 1452–1457.
17. M. Balestri, M. Bober, and W. Bailer, "CDVA experimentation model (CXM) 0.2," ISO/IECJTC1/SC29/WG11/N16274, Geneva, May 2016.
18. G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," In *Int. Conf. Learning Representations (ICLR)*, pp. 1–12, 2016.
19. A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1269–1277.
20. J. Lin *et al.*, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.
21. J. Wang *et al.*, "Learning finegrained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2014, pp. 1386–1393.
22. H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 2167–2175.
23. Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," In *IEEE Int. Conf. Computer Vision (ICCV)*, pp. 814–823, 2017.
24. Y. Bai, F. Gao, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," In *IEEE Int. Conf. Multimedia and Expo (ICME)*, pp. 1452–1457, 2017.
25. G. Serra, C. Grana, M. Manfredi, and R. Cucchiara, "Covariance of covariance features for image classification," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, p. 411.
26. S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 2197–2206.
27. G. Serra, C. Grana, M. Manfredi, and R. Cucchiara, "Gold: Gaussians of local descriptors for image representation," *Comput. Vision Image Understand.*, vol. 134, pp. 22–32, 2015.
28. J. Carreira, C. Rui, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1177–1189, Jun. 2015.
29. T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person reidentification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1363–1372.
30. L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1239–1248.
31. D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person reidentification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 1268–1277.
32. C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 475–491.

33. N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person reidentification," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 858–877.
34. S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2017, pp. 5028–5037.
35. H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2017, pp. 1077–1085.
36. Y. Taigman, M. Yang, M.'A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2014, pp. 1701–1708.
37. Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2014, pp. 1891–1898.
38. O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. Brit. Mach. Vision Conf.*, 2015, pp. 41.1–41.12.
39. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2015, pp. 815–823.
40. J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," arXiv:1506.07310, 2015.
41. C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.
42. Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 499–515.
43. W. Liu, Y. Wen, Z. Yu, and M. Yang, "Largemargin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.
44. W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," In *Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017.
45. J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 4362–4371, 2017.
46. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2016, pp. 779–788.

Lingyu 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 of 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.

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

Shiqi Wang is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. He 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 Peking University, Beijing, China. Contact him at shiqwang@cityu.edu.hk.

Wen Gao is currently a Professor of 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.

Yong Rui is currently the Chief Technology Officer and the Senior Vice President of Lenovo Group. He is responsible for overseeing Lenovo's corporate technical strategy, research and development directions, and Lenovo Research organization, which covers intelligent devices, big data analytics, artificial intelligence, cloud computing, 5G, and smart lifestyle-related technologies. Contact him at yongrui@lenovo.com.

Corresponding author: Ling-Yu Duan