Collaborative Scalable Visual Compression for Human-Centered Videos

Haofeng Huang†, Wenhan Yang†, Wei Xiang‡, Jiaying Liu†, Ling-Yu Duan†
†Peking University, China
‡Bigo, Beijing, China

Abstract—**Machine intelligence systems have been increasingly widely deployed in real-world circumstances, while the conventional human-vision oriented video coding schemes are inefficient to be embedded in large-scale systems and further support a wide range of applications. There have been urgent demands for a new generation of compression framework to efficiently encodes visual data, where the compression and analytics for machine vision and human perception can be jointly optimized. To this end, we propose a novel visual compression framework to provide visual contents with different granularity for both human and machine vision tasks collaboratively. The proposed scalable compression framework maintains the critical semantic information in a basic layer, so that it is capable of supporting the accurate machine vision analysis under a tight bit-rate constraint. It is scalable to provide visual representations of different granularity to support various kinds of tasks, including video reconstruction that serves human vision examination. Experimental results on the human-centered videos have demonstrated the promising functionality of scalable visual coding with improved efficiency for high-performance machine analysis and human perception.**

Index Terms—**Video Coding for Machines, Scalable Visual Compression, Human-Centered Videos**

I. Introduction

With the rise of the new generation of intelligent systems, *i.e.* Smart Cities and Internet of Things (IoT), the bandwidth to support single or multiple machine vision tasks should be taken into consideration when building a distributed intelligent visual system. Existing visual systems rely on video coding technologies originally designed for human vision *e.g.* MPEG-4 AVC/H.264 [1] and High Efficiency Video Coding (HEVC) [2]. With the ever-widening deployment of modern visual intelligence systems, an increasing portion of video content are consumed by machine vision systems, *e.g.*, traffic detection, and action analysis in surveillance videos. Therefore, the efficient compression of video data with compact visual features is expected to facilitate analytic tasks.

To facilitate machine analytics, a branch of visual data coding methods compress the extracted features instead of the original full pictures [3], *e.g.* SIFT [4] for image retrieval, skeleton sequences [5], facial landmark [6], segmentation map [7], depth map [8], and vectorized edge map [9]. However, these methods are usually task-specific, and cannot reconstruct videos to handle a wide range of tasks including human inspection. A possible failure due to the insufficiency of the information may require a costy retransmission to recover. Hence, the procedure of extracting the piece of key information, namely, features, are expected to be *general and scalable*. On one hand, general features are expected to meet common requirements of a group of machine vision tasks; on the other hand, scalable features are expected to connect the video data representation of different granularities, targeting a variety of machine and human vision tasks. Joint performance and efficiency optimization over multiple tasks is actually incurred in developing general and scalable features [10].

Previous visual compression methods handle each task-specific data stream separately when dealing with multiple tasks. They are limited in compression efficiency. There is a lack of connection mechanism to explore the features of different granularity from pixels to semantic features. To maximize the performance of multiple analytics tasks, a generic visual compression method including compressing frame pixels as well as semantic features is crucial for human and machine visions. Hence, joint compression and analytics need to bridge the gap between low-level classical pixel-level redundancy removal and high-level task-specific feature extraction.

To this end, we propose properties for a desirable coding approach: 1) extracting and compressing compact features that maintain the crucial semantic information, 2) providing coarse-to-fine information for different levels of human or machine vision multiple tasks, 3) optimizing features in multiple tasks jointly and collaboratively. This is also aligned with the collaborative compression and analytics paradigm in MPEG standardization effort Video Coding for Machine (VCM) [11].

To efficiently facilitate the emerging distributed intelligent systems, where the visual content can be consumed by both machines and humans, we propose the semantic laddering framework for machine-human collaborative coding. On one hand, the stream of the abstract features is light-weighted and low-cost to support efficient video analytics. A tight constraint on the bit-rate does not hinder it from preserving the critical semantics of the compressed representation. On the other hand, the visual representation can be enriched in a scalable way with fine-grained features.

As an initial attempt, we follow a lot of prior researches to focus on human analytics, as these videos are closely related to our daily life, and human actions convey rich information. In this work, we propose to jointly optimize the compression and

This work is supported by the National Key Research and Development Program of China under Grant No. 2018AAA0102702, the National Natural Science Foundation of China under Contract No.62172020, State Key Laboratory of Media Convergence Production Technology and Systems, and Key Laboratory of Science, Techonology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

978-1-6654-8485-5/22/$31.00 ©2022 IEEE

analytics for human action videos, and figure out a scalable approach for multiple analytics tasks of different granularities with minimal bit-rates. To achieve the scalability in bit-rates while maintaining the critical semantic information for all bit-rate ranges, we propose the learned laddering compression model to utilize the information in a coarser representation to reconstruct the finer ones.

II. Proposed Method

A. Feature Laddering Framework

As illustrated in Fig. 1, we explore the possibility of machine-human collaborative visual compression on human-centered videos, by proposing the feature laddering framework that provides visual information in different granularities.

To preserve the vital semantic information compactly in the videos, and at the same time provide the capacity to reconstruct the full frames, the encoder should extract different levels of representation. The framework is designed with three layers. The basic layer maintains the most critical semantic information in the visual content. The enrichment layer provides the structural information about the visual content, and it can be propagated through time with the basic layer to model dynamics. The basic layer and the enrichment layer combine to provide the middle-level information for spatio-temporal machine analytics. Finally, the visual layer encodes the video based on the information in the first two layers and reconstructs the visual content for human vision examination.

The compression framework is designed to benefit a multi-task machine vision system from the following aspects:

- The basic layer preserves the vital semantic information, so that even when the bit-rates are strictly limited, machine analysis performance is still maintained.
- For down-stream tasks that do not require high-entropy redundant information, the proposed framework can largely reduce the bit-rate consumption and therefore improve the efficiency.
- The framework is scalable in bit-rates as it maintains the capability to reconstruct the visual content based on the already signaled information. Thus, manual examination is also supported with efficiency.

For human-centered videos, we extract the pose as the basic layer representation that keeps the most fundamental semantic information. Poses are, at the same time, compact and light-weight, thus it can be transmitted with high efficiency. The basic layer representation is compressed with a lossless compression scheme. It provides the fundamental features that power a series of high-level action understanding tasks, with high efficiency and low bit-rate consumption.

While poses facilitate many down-stream analytics, some further in-depth machine vision analytics require pixel-level information, *e.g.* semantic segmentation and human parsing [12]. Thus, in the enrichment layer of the laddering framework, the key frames are signaled to the decoder with an intra-frame compression scheme. We observe that pixel-level semantic information can be efficiently propagated from the key frames to the rests with the guidance of the basic layer representations, and it benefits the mid-level analysis.

Finally, the framework is designed to be capable of reconstructing the pixels for human vision. This is important as in most applications, the machine vision algorithms can not provide absolutely confident predictions, and human intervention is needed in the last round. To reduce the bit-rates in video encoding, a compression model that utilizes the already-encoded representations in previous layers is developed.

B. Learned Scalable Visual Compression

To meet the goal of the feature laddering framework, we design the Alpha-Beta Flow model for human-centered video compression. We follow the learned video compression paradigm in existing literature [13], [14] and further make it end-to-end trainable inspired by the scale-space flow model [15]. However, it is not an ideal solution to directly apply the scale-space flow model in this circumstance. The actions of human bodies in human-centered videos are more random than the common motions in other natural videos, *e.g.* camera motion, affine motion. The scale-space flow model conducts the prediction with the following sampling function,

$$\begin{aligned} &\mathbf{x}' := \text{Scale-Space-Warp}(\mathbf{x}, g), \\ \textit{s.t.}\ &\mathbf{x}'[x,y] = \mathbf{X}[x+g_x[x,y], y+g_y[x,y], g_z[x,y]] \end{aligned} \tag{1}$$

where the prediction $\mathbf{x}$ is sampled via flow g from a stack of reference frames $\mathbf{X}$. Each frame in the stack is a differently blurred version of the previously reconstructed frame. It works well for natural motions, while it may face the problem of missing information when an object suddenly moves in or moves out, which is common in human action videos, as the actions are more random. Thus, we modify sampling function with the proposed Alpha-Beta flow model, formulated as,

$$\begin{aligned} &\mathbf{x}' := \text{Alpha-Beta-Warp}(\mathbf{x}, g, \alpha, \beta), \\ \textit{s.t.}\ &\mathbf{x}'[x,y] = \mathbf{X}[x+g_x[x,y], y+g_y[x,y]] \cdot \alpha + \beta, \\ &g, \alpha, \beta = \mathcal{F}(\mathbf{x}, \mathbf{y}, p_{\mathbf{x}}, p_{\mathbf{y}}; \theta_{\mathcal{F}}), \end{aligned} \tag{2}$$

where we generate a flow map g, coefficients α and β with a parametric function $\mathcal{F}$. The Alpha-Beta flow change the sampling hyper space from the original scale-space manifold, *i.e.* stack of blurred images, to the intensity-space manifold, where the parametric function $\mathcal{F}$ decides how confident it is to predict the pixel intensity at the current position by a space flow, and it tunes α to show the confidence. For area that an object pops out or fade away, the model lower its confidence to allow the static prediction β to compensate for the prediction.

With the Alpha-Beta Flow model, we build up the learned video compression scheme to make up the *enrichment layer* and the *visual layer* in the laddering framework. The overall architecture of the model is shown in Fig. 2. The hyperprior based encoder follows the design in [16]. It generates a bit-stream B_I and the reconstructed frame decoded from the bit-stream. The enrichment layer is implemented with the intra encoder, where each frame is independently compressed and decompressed. The visual layer involves the P-Reference

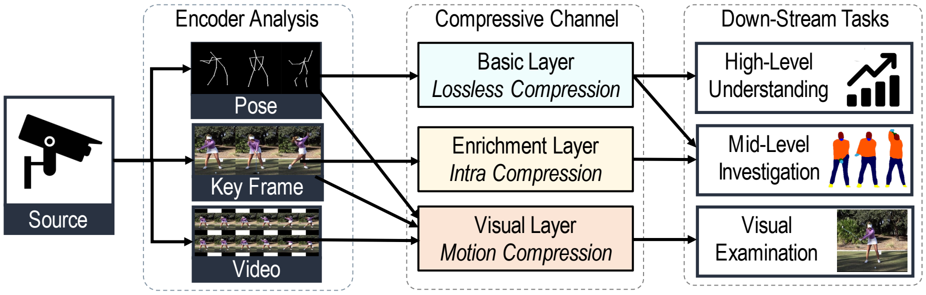

Fig. 1: Feature laddering framework for semantic preserving collaborative visual compression. In applications, the encoder conducts a prior analysis after the video is captured, and it encodes the representations into three layers. The clients request bit-streams with different granularities on their demands, and decode the bit-stream to support various down-stream tasks.

encoder with the Alpha-Beta flow model. It includes a flow encoder to generate the Alpha-Beta flow and the corresponding bit-stream B_F, and consequently makes the predictions. The pose-guided Alpha-Beta flow encoder takes two streams of inputs. The first stream is composed of the reference frame (previously encoded and decoded), the reference pose (embeded in the basic layer of the bit-stream), the target frame (to-be-encoded) and the target pose (also embeded in the basic layer). As the target frame has been embeded in this stream, it requires some bit-rates to signal the information. Meanwhile, the other stream contains only the already decoded representations, *i.e.* the reference frame, reference pose and the target pose. Thus, we do not quantize the latent representation extracted from the second stream and no extra bits should be encoded. The two streams are concatenated in the decoder to finally generate the Alpha-Beta flow for the prediction. We further include a residual encoder here to compensate for the prediction and generates the P-reconstructions with a compensating residual bit-stream B_C.

The pose-guided video compression model is end-to-end trained with the rate-distortion optimization, as,

$$\arg\min_{\theta_I,\theta_F,\theta_R} R + \lambda D, \tag{3}$$

$$R = R_{I_0} + \sum_{i=1}^{N-1} \left(R_{F_i} + R_{C_i}\right), D = \sum_{i=0}^{N-1} d\left(\hat{\mathbf{x}}_i, \mathbf{x}_i\right),$$

where $\theta_I, \theta_F, \theta_R$ correspond to trainable parameters for the intra coder, flow coder and the residual coder, respectively. R denotes the information entropy of the latent representations and D measures the distortion *w.r.t.* original frames and the reconstructions. In the experiments we utilize Mean Squared Error (MSE) as the distortion function d.

C. Supporting Down-Stream Tasks

In this work, we explore two representative down-stream machine vision tasks based on the proposed framework, *i.e.* action recognition and human parsing. For action recognition, the losslessly compressed pose sequences, *a.k.a.* the basic layer, is utilized to conduct the analysis. To conduct accurate parsing, we adopt an image-to-image translation framework with the ResNet backbone consisting of 9 blocks.

To generate the parsing result for a specific frame i, the reference input to the network is organized as the concatenation of the enrichment layer representations, *i.e.* the reconstructed key frame $\hat{\mathbf{x}}_{i-k}$ that corresponds to the frame existing at the previous k-th time step, the corresponding parsing result $\hat{y}_{i-k}$, and the corresponding pose $\hat{p}_{i-k}$. Under the condition of the reference input, the network conducts an image-to-image translation, from the current pose $\hat{p}_i$ to the target output $\hat{y}_i$. This can be formulated as,

$$\hat{y}_i = \mathcal{F}(\hat{p}_i|\hat{\mathbf{x}}_{i-k}, \hat{y}_{i-k}, \hat{p}_{i-k}; \theta), \tag{4}$$

where θ corresponds to the trainable parameters. The parameters are trained with the pixel-wise cross-entropy loss as,

$$\arg\min_{\theta} \sum_{1\leq i\leq N, 1\leq j\leq M} \mathcal{L}_{CE}(\mathbf{y}_{i,j}, \hat{\mathbf{y}}_{i,j}), \tag{5}$$

where $\mathbf{y}_{i,j}$ is the one-hot encoding of the ground truth parsing label at the position (i,j) for a frame of $N \times M$ resolution, while $\hat{\mathbf{y}}_{i,j}$ is the prediction by the translation network.

As not all the pixels are transmitted, the proposed method largely reduces the bit-rate consumption for producing accu-

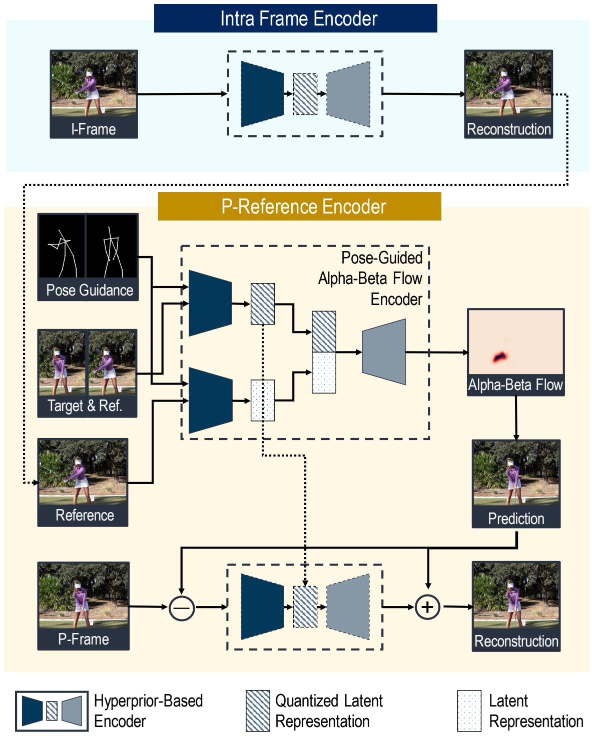

Fig. 2: Pose-guided video compression model based on an Alpha-Beta flow encoder.

TABLE I: Action recognition accuracy on PKU-MMD. *Original* refers to the frames before further lossy compression. *HEVC* stands for frames downsampled to 64×64, and encoded by HEVC (QP=51). We count the bytes for the bit-streams and average it over the frames to show the bit-rates.

Training	Testing	Acc.	Bytes per Frame
Original	*Original*	80.11%	124.21 KB
HEVC *Original* *Original + HEVC*	*HEVC* *HEVC* *HEVC*	26.14% 11.93% 25.57%	39.22 B
Skeleton	Skeleton	75.11%	6.623 B

TABLE II: Evaluation of human parsing on PKU-MMD. *HEVC* refers to parsing results extracted from the frames encoded with HEVC (QP=51). 1^{st} *Frame* (respectively *I-Frames*) refers to the parsing maps extracted from the poses and the first frame (respectively the nearest key frame).

Metric	Labels	*HEVC*	1^{st} *Frame*	*I-Frames*
Acc.	*Average*	52.32%	57.94%	**63.17%**
IoU	*Average*	37.49%	50.93%	**54.91%**
Bytes per Frame		240.19	24.29	121.40

rate action recognition results and the parsing maps, thus to support high-efficiency machine intelligence systems.

III. Experimental Results

A. Experimental Settings

In the experiments, we evaluate the capability of the framework to power machine vision tasks and human visual examination at different ranges of bit-rates. We conduct the experiments on high-quality human-centered video datasets, *i.e.* PKU Multi-Modal Dataset (PKU-MMD) [17].

B. Machine Vision Analysis

1) Action Recognition: To show the general potential performance that the framework can deliver on machine vision tasks, which usually requires the preservation of high-level semantics, we evaluate the action recognition algorithms on the compressed representation of the videos.

For benchmarking, we adopt the algorithm settings in [17], where we apply the Temporal Segment Network (TSN) [18] for action recognition and we adopt the three-layer bi-directional LSTM network for action recognition on skeletons. The comparison of the performances is shown in Table I. Though with the original frames, the down-stream action recognition algorithm achieves the highest performance, it requires a large bandwidth to transmit the frames. When the bandwidth is highly constrained, transmitting the original frames becomes difficult. Compared to HEVC, the proposed framework signals compressed skeleton sequences as the basic-layer representation, and it achieves high performance in action recognition with a much lower bit-rate.

2) Human Parsing: There are some machine vision tasks relying on finer information about the video. In this experiment, we show that the proposed method also facilitate such kind of machine vision algorithms. We evaluate the human parsing performance on the compressed representation of PKU-MMD. The semantic labels are generated by LIP human parsing algorithm [12]. The results are marked as the ground truth of the parsing results. We then compared the parsing results provided by the proposed framework and HEVC respectively, under the constraints of bandwidth.

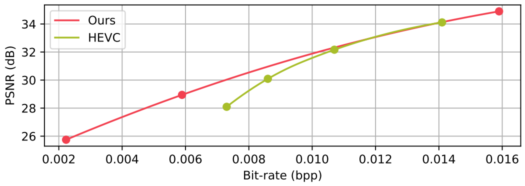

Fig. 3: R-D performance evaluation on PKU-MMD.

The parsing results are generated by the image translation framework. The benchmarking results are shown in Table II. As shown, while HEVC consumes more bit-rates than the proposed method, it does not provide better down-stream parsing results. Utilizing a denser sampling of the reference frames further improves the prediction results while still keeping the bit-rates low.

C. Human Vision Analysis

In this experiment, we evaluate the rate-distortion efficiency of the proposed model, when human examination is required. We train multiple models with the proposed network with different values of λ, to achieve different ranges of bit-rates. Specifically, for the evaluation on PKU-MMD, we select $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}\}$. We set the number of frames between two I-Frames (*a.k.a.*, Group of Pictures, GOP) to 16 in the evaluations for both the proposed model and HEVC, and we adopt the *Low-Delay-P* coding configurations. We evaluate video-level PSNR on the reconstructed frames to compare visual fidelity. For a video sequence, the MSE for all the pixels, with all three channels (*i.e.* R, G, B), in all the frames, is calculated and we compute PSNR from the MSE value. We average the bit-per-pixel (bpp) and PSNR over all the videos, and the rate-distortion curve is shown in Fig. 3. The proposed method achieves better reconstruction quality than HEVC in human examination tasks, and the gain in quality is more significant at lower ranges of bit-rate, where the pose sequences provide useful guidance information.

IV. Conclusion

In this paper, we consider the coding efficiency problem with the bandwidth constraint in machine intelligence systems, and we find that a wide variety of machine vision tasks can be supported efficiently solely by very compact feature representations. Existing visual coding schemes might not maintain the semantic information well at low bit-rates, so they are inefficent for machine vision systems. To address the problem, we propose a laddering framework for machine-human collaborative coding to well support high-accuracy machine vision tasks with lower bit-rate consumption. The proposed model can also be scaled up to meet the need of human examination, and reduce the cost of re-transmission in a machine vision system.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuit System for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Transactions on Circuit System for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] W. Yang, H. Huang, Y. Hu, L.-Y. Duan, and J. Liu, "Video coding for machine: Compact visual representation compression for intelligent collaborative analytics," *arXiv eprints, arXiv:2110.09241*, 2021.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *Proc. of International Conference on Computer Vision*, 2019.

[6] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *arXiv e-prints, arXiv:1805.05563*, 2018.

[7] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] S. Yang, Y. Hu, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: Scalable face image coding," *IEEE Transactions on Multimedia*, vol. 23, pp. 2957–2971, 2021.

[10] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] L.-Y. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[12] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[13] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Z. Chen, T. He, X. Jin, and F. Wu, "Learning for video compression," *IEEE Transactions on Circuit System for Video Technology*, vol. 30, no. 2, pp. 566–576, 2019.

[15] E. Agustsson, D. Minnen, N. Johnston, J. Balle, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[16] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. of International Conference on Learning Representations*, 2018.

[17] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–24, 2019.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. of European Conference on Computer Vision*, 2016.