Collaborative Processing System for Networked Video Applications

Ming Lu[†], Ming Cheng[‡], Qiu Shen[†], Yiling Xu[‡] and Zhan Ma[†]
[†]Nanjing University, [‡]Shanghai Jiaotong University

ABSTRACT

Bandwidth is expensive for networked video at its native spatial resolution (e.g., 4K or even higher). We propose a *collaborative video processing* (CVP) framework where a video stream coded at the lower spatial resolution is delivered for bandwidth saving and upscaled at client with the identical quality of experience (QoE) as the video at its native spatial resolution. Experiments have demonstrated the efficiency of our proposed system for networked video applications (e.g., conferencing, cloud gaming). We have achieved 20% - 50% bitrate reduction without quality degradation measured by multi-scale structural similarity (MS-SSIM). In addition, we can further the bitrate reduction (e.g., 10% or even more) by integrating the learned frame rate up-conversion, through the preliminary studies.

1 INTRODUCTION

Networked video applications prevail in our daily life. Higher bit rate of the compressed video comes with the higher quality content perceived by the user. Typically, compressed streams are adapted to lower bit rate to ensure the smooth network delivery for playback without service blackout, but with compromised quality of experience (QoE). As the video codecs develop, more bitrate reduction at the same quality is brought with the penalty of high computational cost. Moreover, the promotion of a new coding standard usually takes time. *Can we further the coding efficiency on top of the video codecs in existing system?* It is a valuable research and has a significant impact on practical networked video applications.

Recently, deep learning has proved its advanced performance in restoration and synthesis of images [1–4]. Many researchers have implemented DNN-based intra coding, block partition and mode prediction methods in the latest video coding test models. Leveraging the advances in DNNs, we introduce the CVP framework described in section 2.

2 SYSTEM ARCHITECTURE

2.1 Overview

The proposed CVP framework is shown in Fig. 1. A typical spatial down-sampling filter is applied to downscale a high resolution (HR) video input to a low resolution (LR) alternative (e.g., 1080p@60Hz to 960x540@30Hz as exemplified in Fig. 1). Conventional end-to-end video system (highlighted in a dotted-line box) is then utilized to enforce the general compatibility to existing ecosystems. LR video is then upscaled before finally being rendered to the display. Note that we use learned super resolution (SR) and frame interpolation to perform the spatial and temporal resolution upscaling.

2.2 Learned Spatial Resolution Scaling (LSRS)

Efficient SR networks with less parameters are more practical on mobile devices. We select [4] for the time being and extend it to support RGB inputs with five layers in total. We also discuss the impacts of the spatial resolution scaling factor, content dynamics, coding standards, etc, on the performance of the LSRS for the CVP framework.

2.3 Learned Temporal Resolution Scaling (LTRS)

On the top of the aforementioned LSRS, we use a handcrafted method to remove intermediate frames after acquiring the LR frames (e.g., 60Hz to 30Hz). Before the SR operation, a learned frame interpolation is applied to reconstruct the missing frames at its native frame rate. We perform the frame rate up-conversion (or interpolation) by estimating the adaptive convolution kernel [2] to carry out our work.

3 EVALUATION

Four diverse sequences are selected randomly to demonstrate the CVP with only LSRS (as shown in Fig. 2). Distortion is measured using MS-SSIM because it is closer to the perceptual response. Besides, three different spatial resolution scaling factors are investigated in our CVP framework. i.e. LSRS-1.5, LSRS-2 and LSRS-3, corresponding to the scaling factors at 1.5, 2 and 3 respectively. More results are shown in Table 1. In addition to the objective performance, we have also conducted the subjective quality investigation to further evidence the sufficiency of our proposed CVP in the existing end-to-end video systems (as shown in Fig. 3). Moreover, Fig. 4 illustrates the further improvement by integrating the

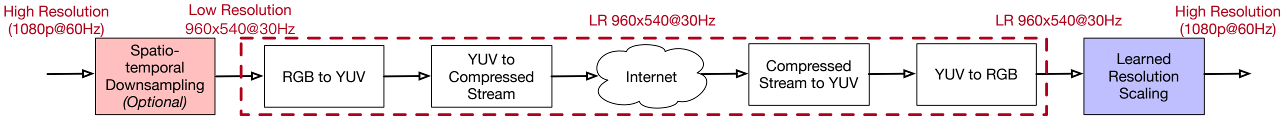

High Resolution (1080p@60Hz)
Low Resolution 960x540@30Hz
LR 960x540@30Hz
LR 960x540@30Hz
High Resolution (1080p@60Hz)
Spatio-temporal Downsampling (Optional)
RGB to YUV
YUV to Compressed Stream
Internet
Compressed Stream to YUV
YUV to RGB
Learned Resolution Scaling

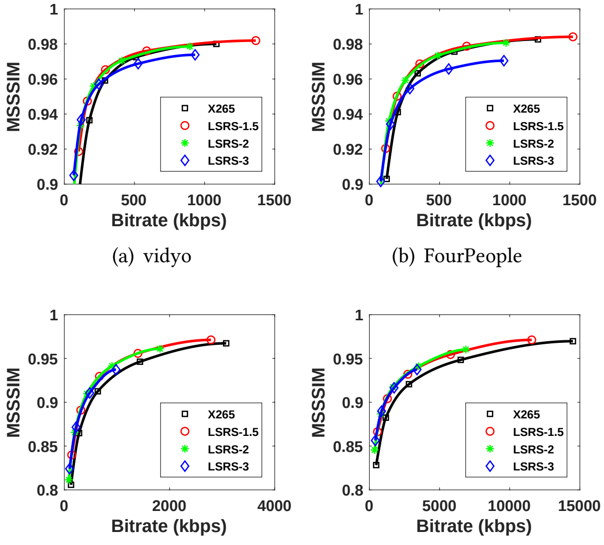
Figure 1: Proposed collaborative video processing via learned resolution scaling

Table 1: BD-Rate Performance Improvement of our CVP

Seq	vidyo	FourPeople	Johnny	LoL	FIFA	WoW#1	WoW#2
BD-Rate†	-	-	-	48.30%	40.85%	40.85%	55.89%
BD-Rate‡	29.9%	20.0%	32.3%	29.36%	34.61%	37.52%	-

†video coded using H.264/AVC; ‡video coded using HEVC.

LTRS on top of the LSRS, with another 10% bitrate reduction at the same QoE.



(a) vidyo

(b) FourPeople

(c) LoL

(d) FIFA

Figure 2: Illustration of BD-Rate efficiency of proposed CVP for videos coded using HEVC

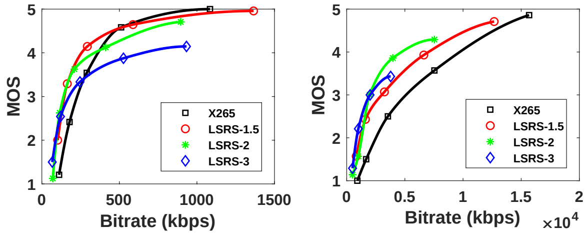


Figure 3: Illustration of MOS versus Bitrate

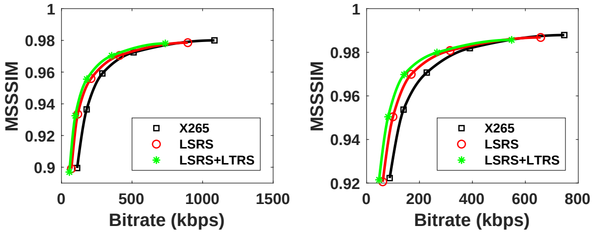


Figure 4: Illustration of BD-Rate efficiency of integrated LSRS + LTRS

4 CONCLUSION

We have developed a *collaborative video processing* (CVP) system for any existing end-to-end video system to stream the video coded at lower resolution, for noticeable bandwidth reduction, and reconstruct the native resolution at the client via learned DNNs. Experimental results have presented 20% - 50% bitrate gain for our proposed LSRS scheme. Another 10% bitrate savings can be achieved by integrating the LTRS.

There are several interesting topics for our further studies, such as the impacts of different DNNs (particularity for those mobile platform favorable networks). In the meantime, how to leverage the federated learning to collect the real-time data to further improve the initial network models is another avenue to explore.

REFERENCES

[1] Chao Dong and et al. 2016. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*. Springer, 391–407.

[2] Simon Niklaus and et al. 2017. Video frame interpolation via adaptive separable convolution. *arXiv preprint arXiv:1708.01692* (2017).

[3] Simon Niklaus and Feng Liu. 2018. Context-aware Synthesis for Video Frame Interpolation. *arXiv preprint arXiv:1803.10967* (2018).

[4] Wenzhe Shi and et al. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1874–1883.