Video Compression by Chroma Prediction Using Semantic Communications

Prabhath Samarathunga
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
prabhath.samarathunga@strath.ac.uk

Yasith Ganearachchi
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
yasith.ganearachchi@strath.ac.uk

Anil Fernando
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
anil.fernando@strath.ac.uk

***Abstract*—Conventional video coding is evolving to meet unprecedented consumer device requirements, but the statistical signal processing based approach may find limitations in handling new media contents. Deep neural network and semantic communication based video compression systems show potential to be used as video encoders and decoders, but reaching the rate distortion performance of state-of-the-art conventional video coding systems remains to be achieved. A novel video compression system by predicting the chroma components of video using the semantically encoded luma component and reference intra-coded frames is proposed and tested against high efficiency video coding (HEVC) for bit rate comparison and rate-distortion performance evaluation. The proposed system demonstrated 18% to 30% saving of bit rate for high and medium motion videos without significant reductions of rate-distortion with the saving increasing with higher group of picture sizes, but low motion videos only demonstrated negligible savings.**

***Index Terms*—Deep Neural Networks, Semantic Communications, Video Compression**

I. Introduction

Video compression is critical to cater the ever growing demand for high definition and high fidelity media consumption. Conventionally, video compression is done using a ever-evolving set of statistical signal processing (SSP) techniques such as those in the H.26X family, which includes advanced video coding (AVC) or H.264, high efficiency video coding (HEVC) or H.265 and the state-of-the-art versatile video coding (VVC) or H.266.

While SSP based video coding standards have kept up with the current demand, future media experiences such as extended reality (XR) and ultra high definition television (UHD-TV) can test the limits of their capabilities. Therefore, interest in deep neural network (DNN) based video coding, as well as emerging concepts such as semantic communications has steadily grown. Semantic communications posits to communicate media content by just sharing the "semantic" or meaning of the content between a sender and receiver who share a common knowledge.

This work proposes a framework to semantically compress a video using luma (Y) component to predict the two chroma components (UV) by applying a semantic communication framework, demonstrating significant bit rate savings over HEVC.

The key contributions of this paper are:

- proposing a semantic communication based system to compress video by predicting UV based on a semantically encoded Y.
- demonstrating improved bit rate savings compared to HEVC.

II. Related Work

Starting with H.261, hybrid video coding using a combination of prediction and transform coding [1] and techniques of prediction, transformation, quantization and entropy coding [2] has been fundamental in the evolution of video coding. The main difference between each generation of video coding being the increasingly complex algorithms used to achieve these aspects [3], and VVC is the current state-of-the-art. However, since implementations of VVC are just entering the market, the fastest growing video coding standard is HEVC, which has a 50% improved bit rate performance over the currently most popular standard - AVC.

However, considering the computational complexity of latest standards, surpassing VVC performance using conventional video compression systems seems highly unlikely. This is leading to more interest in alternative video compression methods, such as AI-based systems, improved perceptual quality measuring systems to optimize compression based on quality, and video coding for machines, which does not require full reconstruction of the videos to be effective in communication [3]. As a response, new video coding frameworks using DNN have also been introduced, such as deep video compression (DVC) which jointly optimized motion and residual information of a video [4].

If the sender and receiver share a prior knowledge, or the context, of the communication the resulting semantic communication system is capable of effectively improving the throughput of a channel significantly [5]. Building on these concepts, use of semantic communications in image and video coding has been examined in [6]–[10]. There have been attempts to construct end-to-end semantic communication systems for images and video using a variety of DNN designs, including hybrids of the traditional models.

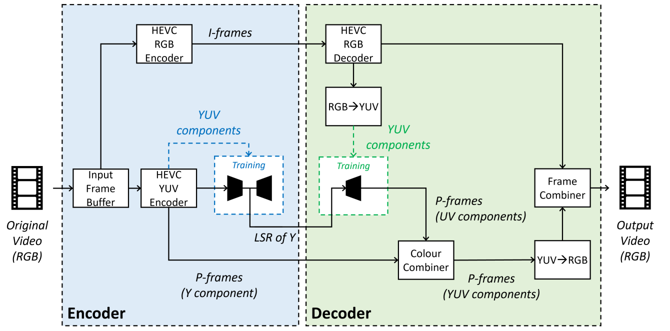


Fig. 1. Architecture of Proposed System

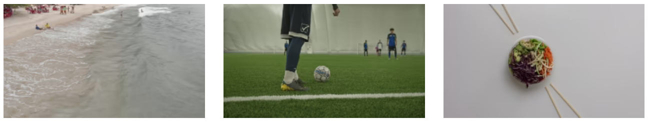
III. Proposed System

The proposed system primarily consists of an autoencoder which is tasked to predict the UV components of each frame of a video encoded in YUV 4:2:0 format based on the Y component. The latent space representation (LSR) created in the process represents a semantically coded version of the input frame which has a very high compressive ratio compared to conventional methods. However, for the LSR to be decoded some context information in the form of intra-coded frames (I-frames) to provide reference to the video sequence and a set of compressed Y samples to provide training at the decoder are needed.

As seen in Fig. 1, the original video is divided to groups of pictures (GOP) and for each group the first I-frame is intra-coded using HEVC in RGB and transferred to the decoder. The frames are also coded using HEVC in YUV colour space, where the encoder-side autoencoder is trained over 100 epochs to minimize the loss between the predicted UV based on Y values and the actual UV values. Once trained, the autoencoder is used to create a LSR for each frames' Y component, which is sent to the decoder as the semantically coded frame.

At the decoder, the semantic decoding network is trained to predict the UV components of the I-frame using the LSR of Y component of the I-frame and the reconstructed UV components of the I-frame from the decoded RGB frame over 50 epochs, and the trained network is used to predict the UV components of the predicted frames (P-frames) based on the LSR of the Y component. Thus, the final P-frame that is reconstructed consists of the received Y component, and the predicted UV components, which is then combined with the I-frame to recover the GOP.

The proposed system was compared against HEVC based on three videos with different motion characteristics: people enjoying the day at the beach (Video 1) [11], people playing soccer (Video 2) [12], and a bowl of avocados and vegetables (Video 3) [13] each with medium spatial information (SI) and medium temporal information (TI), high SI and high TI, and medium SI and low TI respectively, sample frames of which are shown in 2. The bit rate required for each method for GOP sizes of 8, 16 and 32 were calculated and used to compare the bit rate performance of the proposed method compared with HEVC.



(a) (b) (c)

Fig. 2. Three video sequences used in the experiment: (a) Video 1, (b) Video 2 (c), Video 3.

IV. Results and Discussion

The proposed system demonstrated bit rate savings over HEVC in all three videos and across all GOP sizes tested, as seen in Table I where the rates are given in kilo bits per second (kb/s). Significant savings of 20% to 30% were observed in video 1 and video 2, which were relatively high TI, whereas video 3 with low TI only demonstrated slight savings, less than 10%. It is also observed that the bit rate saving increases with the increase of GOP size, with the highest savings in all three videos come when GOP size is 32.

TABLE I
Comparison of bit rates for different GOP sizes

Video	GOP Size	HEVC (kb/s)	Proposed (kb/s)	Saving
Video 1	8	722	589	18%
Video 1	16	664	498	25%
Video 1	32	623	445	28%
Video 2	8	777	618	20%
Video 2	16	731	541	26%
Video 2	32	703	501	29%
Video 3	8	118	117	1%
Video 3	16	77	73	4%
Video 3	32	58	52	10%

However, the improvement of bit rate does not come at the cost of quality, as evident from Table II, where the rate-distortion (RD) performance in terms of peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) show no significant deviation between HEVC and the proposed system. As evident from the results, the proposed system has been able to match RD performance of HEVC in all three video clips.

TABLE II
Comparison of RD performance for different GOP sizes

Video	GOP Size	HEVC PSNR	HEVC SSIM	Proposed PSNR	Proposed SSIM
Video 1	8	30.18	0.9444	30.17	0.9445
Video 1	16	30.14	0.9424	29.98	0.9397
Video 1	32	30.13	0.9412	29.86	0.9381
Video 2	8	31.87	0.9432	30.78	0.9317
Video 2	16	31.67	0.9405	30.09	0.9219
Video 2	32	31.64	0.9389	29.60	0.8941
Video 3	8	28.50	0.9827	28.51	0.9805
Video 3	16	28.49	0.9824	28.48	0.9789
Video 3	32	28.48	0.9821	28.49	0.9796

V. Conclusion

Based on the results, it can be concluded that the proposed system for video compression by chroma prediction using semantic communication is capable to reducing the bit rate required for transfer of high- and medium- motion videos. The bit rate savings in such cases lie between 18% and 30%, with the savings increasing for higher GOP sizes. However, no significant saving can be observed in low motion content videos.

This highlights the potential in chroma prediction based compression using semantic communication concepts, but there is further research required to improve the residual coding performance of the proposed system, as well as for extending the experiments to a larger sample of videos with varying spatial dimensions.

References

[1] B. Bross, K. Andersson, M. Bläser, V. Drugeon, S.-H. Kim, J. Lainema, J. Li, S. Liu, J.-R. Ohm, G. J. Sullivan, and et al., "General video coding technology in responses to the joint call for proposals on video compression with capability beyond hevc," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1226–1240, 2020.

[2] G. Sullivan and T. Wiegand, "Video compression - from concepts to the h.264/avc standard," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 18–31, 2005.

[3] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[4] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[5] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, 2021.

[6] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, 2023.

[7] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, 2023.

[8] M. Zhang, Y. Li, Z. Zhang, G. Zhu, and C. Zhong, "Wireless image transmission with semantic and security awareness," *IEEE Wireless Communications Letters*, 2023.

[9] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, "Wireless deep video semantic transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 214–229, 2022.

[10] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2022.

[11] P. Midtrack. People enjoying the day in a beach. www.pexels.com. [Online]. Available: https://www.pexels.com/video/people-enjoying-the-day-in-a-beach-3150419/

[12] T. Miroshnichenko. People playing soccer. www.pexels.com. [Online]. Available: https://www.pexels.com/video/people-playing-soccer-6077718/

[13] C. of Couple. A bowl of avocados and vegetables. www.pexels.com. [Online]. Available: https://www.pexels.com/video/a-bowl-of-avocados-and-vegetables-7656166/