An Adaptive End-to-end Codec based on Multi-source Fusion for UAV Equipped with 5G NSA Network

Yu Su¹, Qianqian Cheng¹*, Shuijie Wang¹, Yuhe Qiu¹
¹ R&D Department II, China Mobile Chengdu Institute of Research and Development, Chengdu, Sichuan, 610000, China
*Corresponding author’s e-mail: chengqianqian@cmii.chinamobile.com

Abstract. This paper proposed an adaptive end-to-end codec based on multi-source fusion. Our codec adjusts the coding scheme according to the channel condition adaptively, selects deep recursion when the channel is good to obtain sufficient and effective information, and shallow recursion to ensure the transmission of key frames when the channel condition is poor. Besides, communication information is also imported into the codec through parallel branches to avoid confusion caused by multiple sources. We test on the DJI M210 UAV equipped with 5G network to verify the usability of the adaptive end-to-end codec both quantitatively and qualitatively. Under the same target detection level, our codec provides higher compression rate, and adjusts the compression rate to adapt to different conditions.

1. Introduction
The large bandwidth, low delay and high reliability features of 5G have brought great convenience for multi-scene application, such as distributed data collection [1] and high-resolution video transmission [2]. Unmanned Aerial Vehicle (UAV) equipped with 5G has achieved higher transmission efficiency, with a peak download speed of more than 700Mbps and an average download speed of 600Mbps, which is nearly 10 times higher than that of 4G network [3]. UAV has shown great potential in urban management [4], agriculture [5], geology [6] and meteorology [7] et. al.

However, the 5G network has not independent yet, and still relies on the 4G infrastructure. When the network is switched, the output of UAV will fluctuate, resulting in the blocking of high-resolution video transmission [8]. A 5G network uses millimeter wave was established by Verizon [9], maintains the transmission speed of 1Gbps about 300m away from station, but there is still a long way for the millimeter wave to realize extensive coverage.

Video codec plays an important part in video transmission of UAV, which improves the efficiency by coding and decoding. Existing video codec is mainly based on phased artificial designation, such as H.264 [10, 11], H.265(HEVC) [12, 13] et. al. They make good compression through complex interaction between modules, also bring the problem of unified optimization. In 2018, the first end-to-end codec was proposed by Wu, Chao Yuan et al. [14], an integrated network was used to complete the whole coding and decoding process. From another point of view, this is still an independent scheme without consideration of the actual transmission condition.

The adaptive schemes, based on transmission information in 5G network, mainly include adaptive modulation and coding, power control, hybrid automatic retransmission and channel scheduling. Adaptive modulation and coding adjust the mode and rate according to the change of channel. When the quality is good, the modulation level and coding rate are improved to increase throughput. When the channel quality is poor, the modulation level and coding rate are reduced to ensure the transmission...
reliability [15]. Power control reduces the transmit power when the channel is good, and increases the power when the channel is bad [16]. Hybrid automatic retransmission integrates gain at the receiver by adjusting redundancy information and performs retransmission [17]. According to wireless channel measurements, channel scheduling selects better channel for data transmission [18].

The existing 5G transmission information is mainly used in the selection of modulation mode and rate, while video codec aims at ensuring quality of compression. Adaptive codec with multi-source fusion has largely been untouched.

This paper proposed an idea of adjusting the coding scheme according to the channel condition adaptively, selecting deep recursion when the channel is good to obtain sufficient information, and selecting shallow recursion to ensure the transmission of key frames when the channel condition is poor. Based on the work of [14], we establish an end-to-end codec including encoder, quantizer, decoder and entropy coder. Firstly, feature extraction is carried out for key frames, and then the missing frames are generated. Secondly, our codec adds more fine-grained transmission information through parallel branch to fusion multi-source. Transmission information consists of Reference Signal Receive Power (RSRP), Reference Signal Receive Quality (RSRQ) and Signal-to-Noise Ratio (SNR).

We compare our video codec to Chao Yuan Wu’s work [14], and test them on a UAV equipped with 5G NSA. Under the same level of target detection accuracy, our codec provides higher compression rate and can be adjusted to adapt to different communication conditions.

2. Multi-source adaptive video codec

Our codec is an adaptive end-to-end network based on multi-source, which couples communication information with video. By controlling the recursion depth and frame generation, an codec that can be jointly optimized is constructed. We first give an overview on the codec, and then show the detail of each layer of recursive network.

2.1. Hierarchical structure

In the hierarchical structure, when the group size is n, there is a key frame and n-1 reference frames. Take the group size of 12 as an example, the group head element is the key frame. The calculation of the reference frame follows the dichotomy rule. The codec of level 1 generates the 7th frame, and the reference frame is the 1st and 13th frame (the first frame of next group). The codec of level 2 generates the 4th and 10th frames, and the reference frames are [1,7] and [7,13] respectively. Level 3 generates frames [2,3], [5,6], [8,9] and [11,12], which are consistent with the above, and these two frames are symmetrical. When encoding by encoder, the binary stream \( \{I(0), I(1) \ldots\} \rightarrow \{b(0), b(1) \ldots\} \) is generated. After decoding by decoder, the binary stream is restored to frame data \( \{b(0), b(1) \ldots\} \rightarrow \{I(0), I(1) \ldots\} \).

Recursion is implemented by hierarchical structure, and the communication status needs to be evaluated before building the structure. We choose RSRP, RSRQ and SNR to do this evaluation, which are the reflection of actual communication situation. To make it easier, we integrate them into a channel factor \( \alpha \). The calculation process is shown in formula 1, the boundary is obtained by dividing the range of values.

\[
\alpha = \text{sum}(\text{RSRP}, \text{RSRQ}, \text{SNR}) = \begin{cases} 
[\text{-}105] & \text{level1} \\
[\text{-}105, \text{-}94] & \text{level2} \\
[\text{-}94,] & \text{level3}
\end{cases}
\] (1)

With the deepening of recursion level, video gets more frames and provides more complete description. Meanwhile, the hierarchical structure ensures that the coding of each layer is based on the previous layer, making the whole structure more compressible.
As shown in Figure 1, each layer is a network composed of coding, quantization, decoding and entropy coding. Note that the structure of different layers is consistent, and Adam is used to jointly optimize the parameters of multiple steps. The specific encoding and decoding structure will be introduced in Section 2.2.

After convolution in the encoding and decoding part, each pixel in a frame is still treated equally, and redundancy can be further reduced by entropy coding. In this paper, our codec uses the structure distribution probability proposed by Mentzer et al. [19] to map the features and complete the entropy coding part.

2.2. Interpolation with communication information
As shown in Figure 2, our multi-source codec can be divided into four parts: motion estimation (E), context extraction (C), communication fusion (F) and interpolation (P).

Motion estimation reduces the uncertainty of frame insertion and is a key step to improve performance, the relevant test is shown in Section 3. As shown in Figure 2(a), motion estimation is calculated before the network is trained by block motion estimation [20], and sent to the network with training frames. $V_{motion}$ are calculated by formula (2), $V$ means vector, the estimation only exists between context frames.

$$V_{motion} = E(\text{Frame}_{before}, \text{Frame}, \text{Frame}_{after}) \tag{2}$$

Context extraction is based on the Unet proposed by Ronneberger et al [21], which is used to extract the feature maps of context frame. The frame is down sampled firstly, and then up sampled step by step to the size of the original image. Context extraction is independent of the main network, and it is imported into the backbone by parallel branches after extraction.

$$V_{context} = C(\text{Frame}_{before}, \text{Frame}, \text{Frame}_{after}) \tag{3}$$
Communication fusion is the main challenge for the codec proposed in this paper. The structure and change rules of communication data and image are quite different. Based on the inception [22] proposed in GoogleNet, the communication data is also processed as a parallel branch separately. Then, the fusion part imported it into the backbone to avoid the ambiguity caused by the multi-source. Comparisons of different structures for multi-source fusion are presented in Section 3, $V_{fusion}$ is calculated by formula (4).

$$V_{fusion} = F(RSRP, RSRQ, SNR)$$  \hspace{1cm} (4)

After the completion of the above parts, the input includes the motion information obtained by motion estimation, the context feature maps obtained by context extraction, and the channel status provided by communication fusion. The generated frame is calculated by formula (5).

$$\hat{I} = P(I, V_{motion}, V_{context}, V_{fusion})$$  \hspace{1cm} (5)

In the above four parts, motion estimation is an independent module, C, F and I are jointly trained, as shown in formula (6), norm is used as the loss to optimize the whole network.

$$l(\hat{I}, I) = ||\hat{I} - I||_1$$  \hspace{1cm} (6)

3. Experiment

In this part, testing of the adaptive codec based on communication information are implemented, both quantitative and qualitative evaluation are presented.

3.1. Experiment methodology

**Datasets.** Our video datasets are all collected from DJI M210 UAV (as shown in Figure 3. (a)), the resolution is 1920*1080. The 5G signals are provided by China Mobile (as shown in Figure 3(b)), which is works in NSA mode. The dataset contains 2M frames in 53K videos, 45K for training, 4K for validation and 4K for testing. The group size used here is 12. In this part, we test the performance of the proposed adaptive end-to-end network under network fluctuation.

![Figure 3. (a) DJI M210 UAV (b) 5G communication module](image)

**Indication.** Our codec is tested from three aspects: quantity, quality and usability. Bits per pixel (BPP) is used to evaluate the quantity. Multi-Scale Structural SIMilarity (MS-SSIM) and Peak Signal to Noise Ratio (PSNR) are used to evaluate the quality. IOU and mAP of the decoded frames are used to evaluate the usability of our codec.

3.2. Comparisons and analysis

We first test the model locally to verify its usability.

![Figure 4. Performance of codec with different motion estimation](image)
Motion estimation has greatly improved the overall performance of codec, which brings more priori information to the model and reduces the larger ambiguities in frame insertion. In this paper, motion estimation is further evaluated. The video captured by UAV will shift the whole picture with the camera moving. However, the performance of optical flow motion estimation is not satisfactory due to the influence of light. Large area motion estimation results make the reconstructed frames blur and overlap. As shown in Figure 4, the results of optical flow estimation fluctuate greatly, and the non-normalized optical flow estimation brings unstable factors such as the whole picture drift and blur. Therefore, the block motion estimation method used in H.264 is used in the motion estimation part of this paper, it shows a more stable description accuracy in the test.

The integration of communication information is the main challenge of codec design. Firstly, we try to fuse communication information with frame data by adding channels, and share convolution kernel in optimization, but the test results are worse. The change of adjacent frames in video has a completely different mode from the communication information’s. The shared convolution kernel brings great instability to the network, and the generated video frame shows obvious distortion and inconsistent with the ground truth. Finally, based on the idea of inception [22], communication information is added to the network in the way of parallel branches to assist the generation of frames. As shown in Figure 5, both structural similarity and PSNR have been greatly improved in terms of video quality.

Next, the whole codec is tested to verify the applicable scenarios.

As shown in Fig. 6, image compression performs worst. This is not surprising because the time dependence is not taken into account, and the performance of image compression degrades seriously with the decreasing bit rate.

As shown in Figure 6. (a), communication information brings great improvement in BPP, but decreases in video quality at the same time. In fact, it is expected that codec will reduce the depth of network recursion when communication is poor, thus affecting the continuity of video frames. But on the other hand, it is not wise to over infer the unknown frame when the channel is full of interference. Compared with the blocked channel, UAV needs more effective transmission results.

This part evaluates the usability of codec from the aspect of target detection. As shown in Table 1, the ambiguates caused by excessive reasoning is stopped in time, reliable frame data retained does not affect the target detection accuracy. It should be noted that the codec proposed in this paper is not suitable for live broadcast scenes, missing frames may affect the visual experience.
Table 1. IOU and mAP of different codec

| Codec                        | IOU | mAP |
|------------------------------|-----|-----|
| Video codec [14]             | 0.85| 82  |
| Video codec+cominfo (ours)   | 0.80| 78  |

4. Conclusion
This paper presents an adaptive end-to-end video codec based on multi-source. The main idea is to adjust the coding scheme according to the channel condition adaptively. When the channel condition is good, deep recursion is applied to obtain sufficient and effective information. When the channel condition is poor, shallow recursion is applied to ensure the transmission of key frames, so as to realize the accurate and fast adaptive.

The codec proposed in this paper has been greatly improved in BPP, but this does not mean that we have found an all-round solution suitable for UAV video transmission. Although the improvement here does not affect the accuracy of target detection, there is still a big gap compared with the real-time live broadcast which requires high visual experience. At the same time, it has great optimization space in engineering.

The codec proposed in this paper verifies the possibility of adaptive end-to-end codec based on multi-source fusion, and gives its superior performance in adaptive adjustment.

Acknowledgments
We would like to thank Yuanlin Xu, Ziyang Li, Yike Ren and Hangyu Li in UAV data collection. This work is supported by CMLC team of China Mobile Chengdu Institute of Research and Development.

References
[1] A. Ogawa et al., "Field Experiments on Sensor Data Transmission for 5G-Based Vehicle-Infrastructure Cooperation," 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 2018, pp. 1-5.
[2] T. Higuchi et al., "Video Sending Rate Prediction Based on Communication Logging Database for 5G HetNet," 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018, pp. 1-6.
[3] B. Li, Z. Fei and Y. Zhang, "UAV Communications for 5G and Beyond: Recent Advances and Future Trends," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2241-2263, April 2019.
[4] X. Hu, B. Pang, F. Dai and K. H. Low, "Risk Assessment Model for UAV Cost-Effective Path Planning in Urban Environments," in IEEE Access, vol. 8, pp. 150162-150173, 2020.
[5] P. Tokekar, J. V. Hook, D. Mulla and V. Isler, "Sensor Planning for a Symbiotic UAV and UGV System for Precision Agriculture," in IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1498-1511, Dec. 2016.
[6] N. Chen, N. Ni, P. Kapp, J. Chen, A. Xiao and H. Li, "Structural Analysis of the Hero Range in the Qaidam Basin, Northwestern China, Using Integrated UAV, Terrestrial LiDAR, Landsat 8, and 3-D Seismic Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 9, pp. 4581-4591, Sept. 2015.
[7] L. Wirth, P. Oettershagen, J. Ambühl and R. Siegwart, "Meteorological path planning using dynamic programming for a solar-powered UAV," 2015 IEEE Aerospace Conference, Big Sky, MT, 2015, pp. 1-11.
[8] Q. Yuan, Q. Qian, Y. Mo and H. Chen, "Research on Mixed Planning Method of 5G and LTE," 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 489-493.
[9] https://www.verizon.com/about/our-company/5g/what-millimeter-wave-technology
[10] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H.264/AVC video coding standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol.
13, no. 7, pp. 560-576, July 2003.

[11] H. Schwarz, D. Marpe and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, pp. 1103-1120, Sept. 2007.

[12] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[13] J. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC)," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1669-1684, Dec. 2012.

[14] Wu, Chao Yuan, N. Singhal, and Krhenbühl, Philipp, "Video Compression through Image Interpolation." European Conference on Computer Vision Springer, Cham, 2018.

[15] A. Tsakmalis, S. Chatzinotas and B. Ottersten, "Modulation and Coding Classification for Adaptive Power Control in 5G Cognitive Communications," 2014 IEEE 15th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Toronto, ON, 2014, pp. 234-238.

[16] I. O. Sanusi, K. M. Nasr and K. Moessner, "Channel Assignment and Power Control for D2D-Enabled Cellular Networks," 2019 International Conference on Computing, Electronics & Communications Engineering (iCCECE), London, United Kingdom, 2019, pp. 225-228.

[17] J. Yeo, S. Park, J. Oh, Y. Kim and J. Lee, "Partial Retransmission Scheme for HARQ Enhancement in 5G Wireless Communications," 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 2017, pp. 1-5.

[18] B. Wang, Y. Zhu and J. Kang, "Two Effective Scheduling Schemes for Layered Belief Propagation of 5G LDPC Codes," in IEEE Communications Letters, vol. 24, no. 8, pp. 1683-1686, Aug. 2020.

[19] Oord, Aaron Van Den, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel Recurrent Neural Networks." (2016).

[20] Richardson, Iain E. G. "Video Codec Design: Developing Image and Video Compression Systems." John Wiley & Sons Inc (2002).

[21] Ronneberger, O., Fischer, P., Brox, T., "U-net: Convolutional networks for biomedical image segmentation". MICCAI (2015)

[22] Szegedy, Christian, et al., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." (2016).