AI-based Video Processing and Its Application to Wireless Transmission

Li YUE*, Naoki MAEDA, Shigeharu TOYODA, Shouhei OGAWA, Kazuya HANDA, and Yasuhito FUJITA

In recent years, AI-based video analysis has been widely used in various fields. Demand for higher definition and wider area viewing has led to the development of cameras with higher resolutions such as 4K and 8K. As a result, the cost of video data transmission, storage, and analysis has increased significantly. In response to this, we have developed a prototype AI-based video processing technology (AVP) that can significantly reduce the amount of video compression data and cloud AI computation cost after transmission. In a field test conducted at a factory, AVP reduced the average bit rate by 92.2% compared to conventional video compression technology, while reducing the AI computation cost on the cloud side, enabling highly accurate analysis of workers' movement flow using high-resolution video.

Keywords: AVP (AI-based video processing), high-efficiency video compression, AI, factory IoT, 5G

1. Introduction

Video has been used just for fun for a long time, but its application expands in recent years as AI-based video analysis of people, things, and events spreads in a wide range of fields. At the same time, in response to the need to see a wider range with higher definition, the use of cameras with high resolution, such as 4K and 8K, is also increasing. As a result, it is expected that a large amount of video data will be transmitted from a large number of high-resolution cameras installed in various devices and places, such as cars, robots, factories, and roads, and the transmission amount of video data is expected to increase explosively in the future.

AI-based video analysis often uses deep learning. Analysis that requires a lot of computing resource is difficult to be handled at the edge (camera side) and therefore the use of cloud computing is an effective solution. However, the problem is that the transmission of a large amount of video data to cloud imposes an excessive load on the network bandwidth. In recent years, the high-speed, large-capacity 5G communication system (fifth generation mobile communication system) has been spreading, but to transmit a large amount of video data, even if using a 5G network, improving transmission efficiency is a major issue. For this reason, the development of not only network technology but also edge processing technology is necessary.[(1)-(3)]

Against this background, we have developed a highly efficient video compression transmission technology called "AVP" (AI-based Video Processing). The video data compressed and transmitted by AVP is assumed to be used for AI analysis or event check by human sight. For this reason, rather than pixel-level image reproduction which conventional video compression technologies have been aiming for, AVP focuses on the AI analysis accuracy and image reproduction at the semantic level after the transmission and decompression. AVP is possible to compress video data to 1/10 or less compared to the conventional compression technology. In this paper, we introduce the features of AVP and show its effects on factory IoT.

2. High-Efficiency Video Compression Transmission Technology "AVP"

AVP is characterized by its ability to transmit necessary video information in real-time while minimizing the amount of compressed data, by mimicking the human eye's characteristic of consuming less energy by looking clearly only at areas of interest and blurring other areas. As shown in Fig. 1, AVP consists of two core process blocks: the attention area extraction AI and the area-wise differentiated video compression. The attention area extraction AI extracts areas in the video that affects the AI analysis accuracy or human understanding of the situation (hereinafter referred to as "attention areas"). In the process of area-wise differentiated video compression, only attention areas maintain high quality, and those of other areas are degraded. This method makes it possible to significantly increase the compression ratio without losing key information.

AVP has the following four features.

(1) By differentiating quality for each area, the bit rate can be reduced to 1/10 or less compared to the conventional technologies that compress the entire image with uniform quality.
(2) The compressed video data can be played back with general-use decoders and video playback software due to the use of standard-compliant compression technologies.
(3) Attention area coordinate information is stored in the vendor extension area of the compressed video data and can be used to reduce the processing load at the receiver side.
(4) Adaptive bitrate control in response to the available network throughput fluctuation.

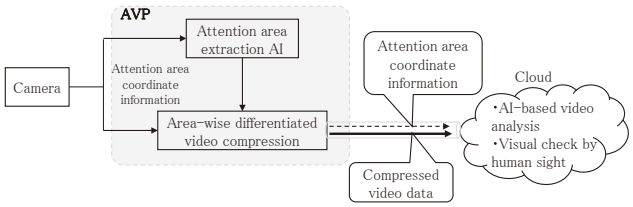


Fig. 1. AVP block diagram

The following are three representative applications of AVP:

(1) The factory IoT field, where there is an increasing need for high-precision AI analysis using high-definition video

(2) Self-driving buses and taxis, which require remote monitoring inside and outside vehicles via mobile communication network that causes high cost for data transmission

(3) Outdoor autonomous robots, such as delivery robots and security robots, which also require remote monitoring and control via mobile communication network.

Details of AVP's functions are described below.

2-1 Attention area extraction AI processing

The attention area extraction AI uses deep learning to select areas in the video that contain objects to be analyzed by the cloud AI (hereinafter referred to as "attention objects") and help human beings understand the situation.

The attention area extraction AI divides an image into multiple blocks, as shown in Fig. 2, and determines whether there is any attention object (vehicles in the case of Fig. 2) in each block. Object detection AI is usually used to identify the positions of attention objects. However, compared to object detection AI, attention area extraction AI is much simpler and can be realized with a lighter-weight neural network because it does not determine the type and accurate position of each object, unlike object detection AI. The attention area extraction AI can determine not only whether there is any attention object in each block, but also whether it is small or large. Hereinafter, blocks including any small attention object are referred to as "fine attention areas," those including middle to large attention objects as "coarse attention areas," and those including no attention objects as "non-attention areas."

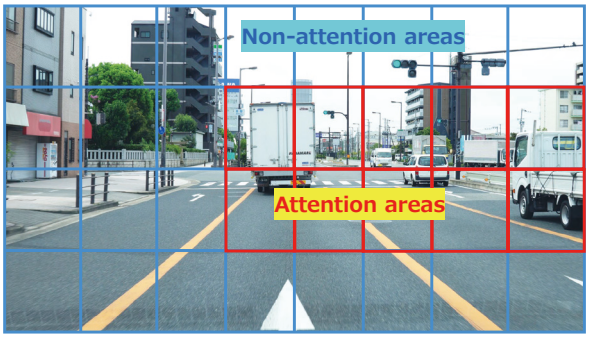


Fig. 2. Output example of attention area extraction AI

2-2 Area-wise differentiated video compression

In the process of area-wise differentiated video compression, fine attention areas are compressed to "high quality" (low compression ratio), coarse attention areas are compressed to "medium quality" (medium compression ratio), and non-attention areas are compressed to "low quality" (high compression ratio) (See Fig. 3). Video quality after compression (compression ratio) is controlled by changing the quantization parameter (QP).*[1]. The QP value of each area is optimized so that the analysis accuracy of the cloud AI does not degrade compare to the uncompressed video. Since AVP uses a standard-compliant compression technology, compressed video can be played back with general-use decoders and video playback software. Therefore, there is no need to prepare a dedicated decoder.

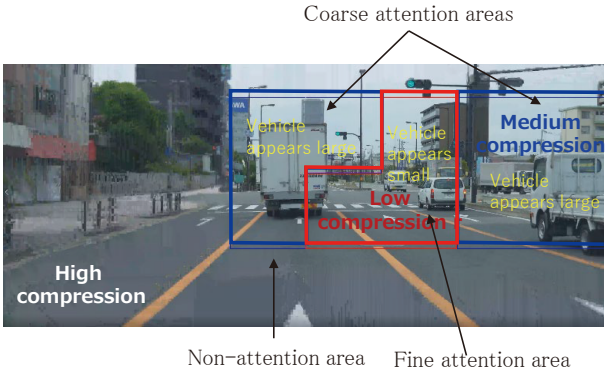


Fig. 3. Example of attention areas and area-wise differentiated video compression

2-3 Reduction of processing load of the cloud AI by using attention area coordinate information

Attention area coordinate information, the output of the attention area extraction AI, can be used to reduce the processing load of the cloud AI. It is known that the amount of computing resources needed by AI for image analysis is generally proportional to the number of pixels of the input image. In the case of AVP, instead of inputting the entire decoded image to the AI, only the attention areas in the image clipped based on attention area coordinate information are input to the AI system, as shown in Fig. 4, which

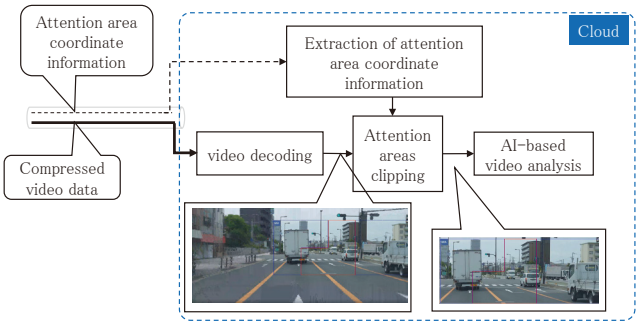


Fig. 4. Diagram showing the mechanism of using attention area coordinate information

can significantly reduce the amount of computation resources.

2-4 Bit rate control based on network condition monitoring

Fluctuation of available network bandwidth*2 is known as a serious issue of remote monitoring and control via mobile communication networks. This fluctuation may cause various problems, such as large changes in the arrival interval of video packets and large disturbances in images due to packet loss.

AVP uses a video transmission protocol that ensures stable communication quality even on unstable transmission paths and enables the distribution of the encrypted video. It is possible to obtain various network statistics from the protocol in real-time and therefore we are working on the development of transmission bit rate control technology that makes use of these statistics. Through detection of the network congestion in real-time by combining multiple statistics, AVP increases the transmission bit rate until it exceeds the available bandwidth, and decreases the transmission bit rate when it detects that the available bandwidth has been exceeded, as shown in Fig. 5.

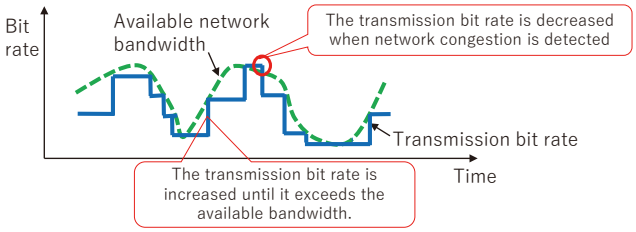


Fig. 5. Bit rate control in response to the detected available bandwidth

3. Application Example of AVP to Factory IoT

We are working with SoftBank Corp. to improve productivity at factories through DX (Digital Transformation) using 5G, AI, and IoT. By analyzing data sent from various sensors, equipment, and cameras in a factory to a cloud server through our industrial 5G terminal, it becomes possible to easily and stably visualize the status of the production site with a simple network system configuration.

In our past demonstration experiments,[4],[5] full HD cameras (1920 × 1080 pixels) were installed in a factory; individual workers in images captured by the cameras were identified using a skeleton recognition AI,*3 and their actual work time was visualized in real-time based on analysis of their flow patterns. This system makes it possible to quickly provide feedback to the staff about tasks that took much longer time than planned. However, due to the three constraints (I. analysis time, II. GPU memory usage for analysis, and III. transmission data volume), full HD video was resized to 640 × 640 pixels before it was processed by the skeleton recognition AI. Due to the low resolution, it was not possible to identify workers who appeared small in the images because their outlines were not clear, and as a result, the recognition rate of the skeleton recognition AI system was as low as 58.2% and much lower than the level required to improve work efficiency, 95%. As a countermeasure, it is possible to perform AI analysis using video with the original resolution without resizing it, or video with even higher resolution, such as 4K and 8K. This, however, significantly increases the amount of data to be transmitted to the cloud server, the AI processing load, and the AI processing time. Regarding the transmission data amount, even if using a high-speed/large-capacity 5G network, improving transmission efficiency is a major issue. In order to address these issues, we configured a system using a 4K camera and an AVP prototype, as shown in Fig. 6, and we were able to significantly improve the recognition rate of the skeleton recognition AI without increasing the 5G network traffic and the processing load on the cloud server.

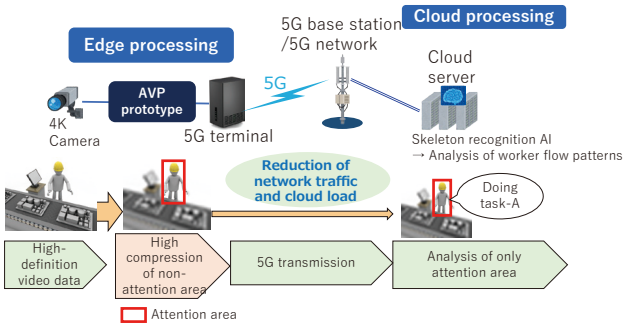


Fig. 6. System configuration after introduction of AVP

When 4K high-definition video was transmitted as it was, a large number of packet losses (data losses) occurred due to insufficient bandwidth even though a 5G network was used. In contrast, when AVP was used, the average bit rate was reduced by 92.2%, and the video data were transmitted without any problem. In the process of image analysis on the cloud, the images of areas including workers were clipped from 4K images based on the attention area coordinate information. Then, the clipped images were resized to 640 × 640 pixels and input to the skeleton recognition AI. Since the clipped images of workers rarely have a resolution significantly higher than 640 × 640, it was able to process them without resizing in most cases and therefore can improve the recognition accuracy. As a result, the worker recognition rate was improved to 99.1%. In addition, since the skeleton recognition AI processes small clipped images, its processing load is independent on the camera's shooting resolution (4K), which makes it possible to perform analysis of higher resolution videos without increasing the computing capacity of the cloud server.

Table 1 shows a comparison before and after the introduction of AVP. By introducing AVP, it was possible to use high-definition cameras, which greatly improved the recognition rate of the skeleton recognition AI. As a result, more accurate work time visualization system was created, which led to the improvement of work efficiency. Furthermore, the lower volume of video data contributed to the reduction of video storage costs.

Table 1. Comparison of before and after AVP introduction

	Before AVP introduction	After AVP introduction
Camera resolution	1920 × 1080	3840 × 2160
Average bit rate	1.6 Mbps	0.5 Mbps
Recognition rate of the skeleton recognition AI	58.2%	99.1%

4. Conclusion

We have developed a high-efficiency video compression and transmission technology "AVP," and confirmed a significant reduction in the transmission bit rate and the processing load of cloud AI through demonstration experiments at a factory. We believe that this technology is also effective for self-driving buses/taxis and outdoor autonomous robots, such as delivery robots and security robots, since they all require remote monitoring using video transmitted via mobile networks. We are planning to confirm the effectiveness of AVP in these applications through demonstration experiments in the future.

5. Acknowledgments

We would like to thank Designated Professor Hiroshi Murase, Associate Professor Daisuke Deguchi, and Lecturer Yasutomo Kawanishi of the Graduate School of Informatics, Nagoya University, for their helpful guidance and advice for this research and development.

Technical Terms

∗1 Quantization parameter (QP): QP is a parameter that defines the quantization step size. For every increase in QP by six, the quantization step size doubles.

∗2 Available network bandwidth: The bandwidth that can be used, which is given by excluding the unavailable bandwidth used by other parties from the physical bandwidth of the bottleneck link in the communication path.

∗3 Skeleton recognition AI: A deep learning algorithm for detecting human joint points.

References

(1) Y. Shinohara, H. Itsumi, B. Florian, T. Iwai, "Video compression estimation recognition accuracy for remote site object detection," IWCMC (2020)

(2) L. Galteri, M. Bertini, L. Seidenari, A. Del Bimbo, "Video Compression for Object Detection Algorithms," ICPR (2018)

(3) Tomonori Kubota, Takanori Nakao, Eiji Yoshida, "A high-compression video coding method for video analysis using Deep Learning," IEICE Technical Report, vol. 119, no. 456, IE2019-121, pp. 121-126 (March 2020)

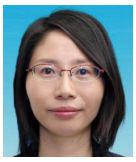
(4) Press release (2019/11/12) by Sumitomo Electric Industries, Ltd. and SoftBank Corp. https://sei.co.jp/company/press/2019/11/prs089.html

(5) Press release (2021/6/9) by Sumitomo Electric and SoftBank Corp. https://sei.co.jp/company/press/2021/06/prs047.html

Contributors

The lead author is indicated by an asterisk (∗).

L. YUE*
• Ph.D.
Group Manager, Information Network R & D Center



N. MAEDA
• Senior Assistant General Manager, Information Network R&D Center

S. TOYODA
• Assistant General Manager, Information Network R&D Center

S. OGAWA
• Assistant General Manager, Information Network R&D Center

K. HANDA
• IoT R&D Center

Y. FUJITA
• Department Manager, IoT R&D Center