Video Data High-Compression Technology Specialized for AI Analysis

–Toward Use in Cloud and 5G–

Tomonori Kubota Takanori Nakao Eiji Yoshida Makoto Kubota

This proprietary technology from Fujitsu was already introduced in a press release published on March 5, 2020 titled "Fujitsu Streamlines AI Video Recognition with High-Quality Compression Technology". This article provides further details of the technology with a focus on its technical aspects.

In recent years, there has been an increase in use of video images in various fields such as detection of suspicious people on the street and analysis of operator behaviors on production lines. With the advent of fifth-generation mobile communications system (5G) services featuring high-speed, large-capacity communications, which enable the use of ultra-high-definition 4K video and high-volume video captured by cameras, there are also high expectations that such images can be analyzed by AI to recognize small behaviors overlooked by the human eye. AI analysis of video data requires a massive number of calculations. As the number of cameras used changes, the number of calculations also changes. The cloud is suitable for these types of calculations, as it is highly scalable and users only need to pay for what they use. Conventionally, however, it has been difficult to use the cloud for this purpose, because transmission of large video data uses up the network bandwidth. To solve the above problem, Fujitsu has independently developed a technology that achieves high compression of video data specialized for AI analysis. The application of the new technology enables video analysis with the cloud. This article describes the problems with conventional video compression technology and provides details of our video data high-compression technology for AI analysis, along with its evaluation results.

1. Introduction

In recent years, the use of video images captured by cameras is increasing in various locations such as on the street and on production lines. The advent of 5th-generation mobile communications system [1] (5G) service in Japan, in particular, is expected to contribute to an explosive increase in the number of 4K ultra-high-definition video images. And, there is expected to be a sharp increase in demand for AI analysis of video data in various business areas such as marketing and manufacturing quality inspections [2-4].

Deep learning is an AI technique that is generally used to automatically analyze high-volume video data without human intervention. Deep learning involves a large number of calculations. The number of calculations also changes, as more cameras are installed to expand the surveillance area, for instance, or the number of monitoring locations changes, depending on each time period. The cloud is highly scalable and suitable for these types of calculations, as users only need to pay for what they use [5]. Conventionally, however, it has been difficult to use the cloud for this purpose, because transmission of high-volume video data from the field to the cloud uses up the network bandwidth.

To solve the above problem, Fujitsu has independently developed a technology that can compress image data much more than the conventional compression technology intended for visual confirmation by humans, by optimizing image data to the minimum

quality that AI can recognize [6, 7]. This reduces the use of network bandwidth, enabling the utilization of the cloud, without compromising recognition accuracy.

This article is structured as follows. First, the article describes the conventional video compression technology as well as its problems. Then, it describes our new technology, which achieves high compression of video images specialized for AI analysis, along with the evaluation results. And, at the end, it provides a summary and mentions Fujitsu's future plans for the technology.

2. Conventional video compression technology and its problems

This chapter describes the conventional video compression technology and problems that may arise when applying it to AI analysis.

2.1. Conventional technology intended for visual confirmation by humans

Video compression technology is used to compress video images to reduce their data sizes so as to reduce the communications cost required to exchange high-quality video images and reduce the storage cost required to store such images. In general, increasing the compression ratio reduces the data size, but it also reduces the image quality. Therefore, it is important to find a balance between the two.

Conventionally, international standards and recommendations for video compression methods (hereafter, standards), such as H.265/HEVC [8], have been used to determine the compression ratio based on image quality intended for visual confirmation by humans. More specifically, the compression ratio is set to achieve image quality that is comfortable for humans, when looking at each of the "blocks" that divide video images frame by frame [9].

2.2. Data size problem with conventional technology

Figure 1 shows the pixels that are focused on, when YOLOv3 [10], an example of AI for image recognition and other video analyses, recognizes or detects (hereafter, recognizes) the positions of humans in a certain image, which are expressed as white dots, using a visualization technique [11]. This result suggests that the pixels focused on by AI for recognition only account for a very small part of the image. Since conventional compression technology also sets blocks that are not required for AI video recognition as high image quality, the data size of the entire video tends to become large, which has been a problem.

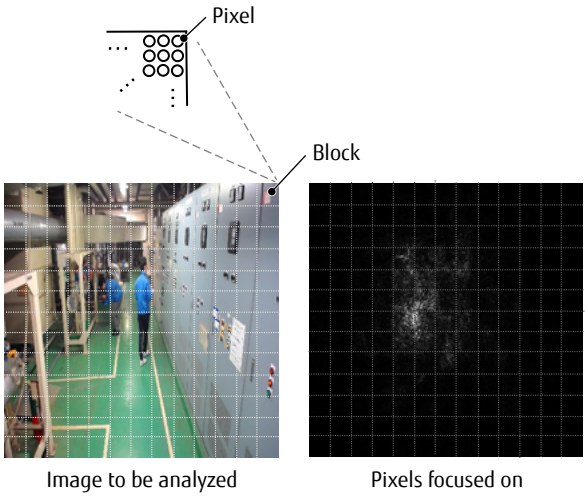


Figure 1
Relationship between pixels focused on by YOLOv3 and blocks.

3. Video data high-compression technology specialized for AI analysis

This chapter describes the technology used to achieve high compression of video data with image quality specialized for AI analysis.

3.1. Approach using this technology

This technology is implemented in two steps.

The first step is to increase the compression ratio of images and try AI object recognition to understand how much the recognition accuracy degrades. Recognition accuracy is a value that is output by AI to indicate how accurate AI itself regards each recognition result. This step is detailed in subsection 3.2.

The next step is to determine the highest compression ratio (critical compression ratio) within the range where AI can recognize objects in the same manner as in the uncompressed state. The compression ratio is determined finely by block to minimize the data size. Finally, each block is compressed at the determined critical compression ratio. This step is detailed in subsection 3.3.

3.2. Understanding the relationship between compression ratio and degradation of AI recognition accuracy

This section describes the three processes used to understand the relationship between the compression ratio and the degradation of recognition accuracy, based on **Figure 2**.

(1) Quantifying the degree of recognition accuracy degradation by compression

First, the entire uncompressed image is compressed at a certain uniform compression ratio, and then decoded. The decoded image is called "image decoded after compression." As some information is lost during the compression, the quality of the image decoded after compression is worse than the quality before the compression. Next, with the uncompressed image and the image decoded after compression as input values, a deep learning model such as YOLOv3 is used to process object recognition and obtain recognition results and recognition accuracy. Finally, the difference in recognition accuracy between the uncompressed image and the image decoded after compression is calculated to quantify the degree of recognition accuracy degradation from the uncompressed image.

The deep learning model outputs a probability value (0 to 1) that indicates how accurate the input image is with respect to the whole pattern (e.g. humans, vehicles) learned as recogni tion targets. This represents the recognition accuracy and the pattern of highest recognition accuracy is regarded as the recognition result. The difference in the recognition accuracy calculated based on the uncompressed image and the image decoded after compression represents the degree of degradation from the uncompressed image.

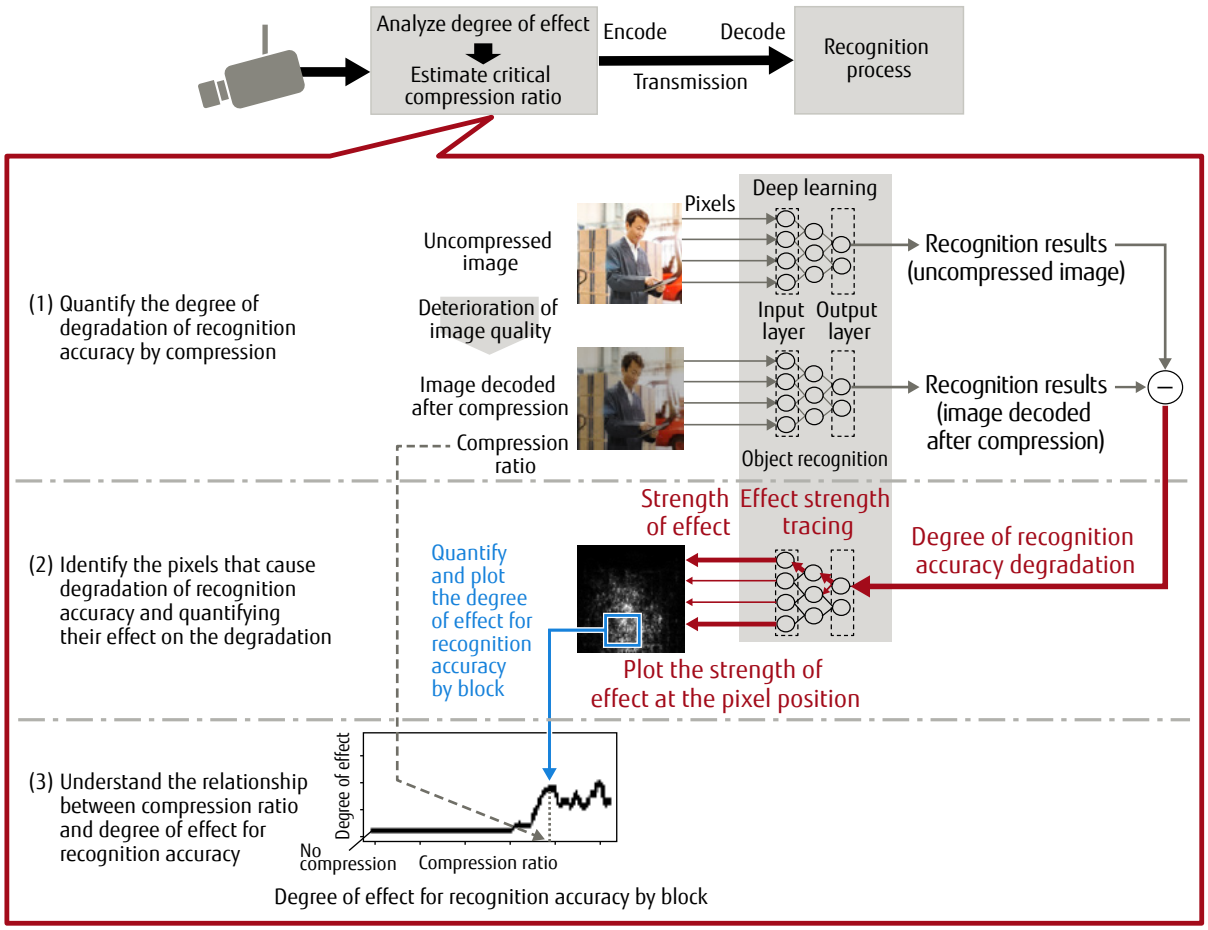


Figure 2
Understanding compression ratio and its impact on AI recognition.

(2) Identifying pixels that cause recognition accuracy degradation and quantifying their effect on degradation

First, the pixels that cause the recognition accuracy to degrade and their effect on recognition accuracy degradation are investigated by using the back-propagation method [12]. More specifically, the degree of recognition accuracy degradation calculated in process (1) is entered in the output layer of the deep learning model and the flow of the object recognition process is traced in the reverse direction. This can obtain the positions of the pixels that cause the degradation of recognition accuracy and the strength of the effect that the pixels have on recognition accuracy degradation (hereafter, degree of effect for recognition accuracy). Next, the degrees of recognition degradation impact by pixel are aggregated by block. This can quantify the degree of effect for recognition accuracy by block.

(3) Understanding the relationship between compression ratio and degree of effect for recognition accuracy

The quantification of the degree of effect for recognition accuracy performed in (1) and (2) is executed by changing the compression ratio for the same image. This makes it possible to identify the relationship between the compression ratio and the degree of effect for recognition accuracy by block, as shown in the chart in Figure 2.

3.3. Determining critical compression ratio and compressing an image

Based on the relationship between the compression ratio and the degree of effect for recognition accuracy calculated in subsection 3.2, this subsection explains how to determine the critical compression ratio by block at which objects can be recognized, and describes the process used to compress each block. **Figure 3** provides an overview of determining the critical compression ratio and compressing an image.

(1) Determining the critical compression ratio by block

Based on the chart obtained in subsection 3.2, the critical compression ratio is set as described below, for the following reasons, if the degree of effect for recognition accuracy increases suddenly at a certain compression ratio when changing the compression ratio.

- The degree of effect for recognition accuracy by block is the minimum in the uncompressed state.
- The block with the largest degree of effect for recognition accuracy contains the most information required for recognition and this means that such information is easily lost through compression.
- If the degree of effect for recognition accuracy increases suddenly at a certain compression ratio, the recognition accuracy also degrades rapidly. Therefore, the compression ratio must not be exceeded.

The process of applying the above standards to determine the critical compression ratio is described in the example of recognizing a person and his face in Figure 3. The compression ratio is changed from low to high in the chart of the compression ratio and the degree of effect for recognition accuracy by block. In this case, the degree of effect for recognition accuracy increases suddenly in blocks near the face of the person at first, and then the degree also increases suddenly in blocks near body parts of the person other than the face. The critical compression ratio for each of these blocks should be lower than the compression ratio at which the degree increases suddenly. On the other hand, the degree of effect for recognition accuracy does not change in background blocks, even when the compression ratio increases. Therefore, the critical compression ratio for these blocks should be set to the maximum possible value.

(2) Compressing each block of the image at the critical compression ratio

Each block of the image is compressed by setting the critical compression ratio calculated in (1), according to the existing standard for setting the compression ratio.

The processes in subsections 3.2 and 3.3 are performed for each image frame. Even if the person moves over time, therefore, the critical compression ratio can be changed for each block in real time according to the position of the person in each frame.

The above technology can achieve high compression of video data, while maintaining the recognition accuracy.

4. Evaluation

This section describes the evaluation of the compression ratio and recognition results of image data using the newly developed technology.

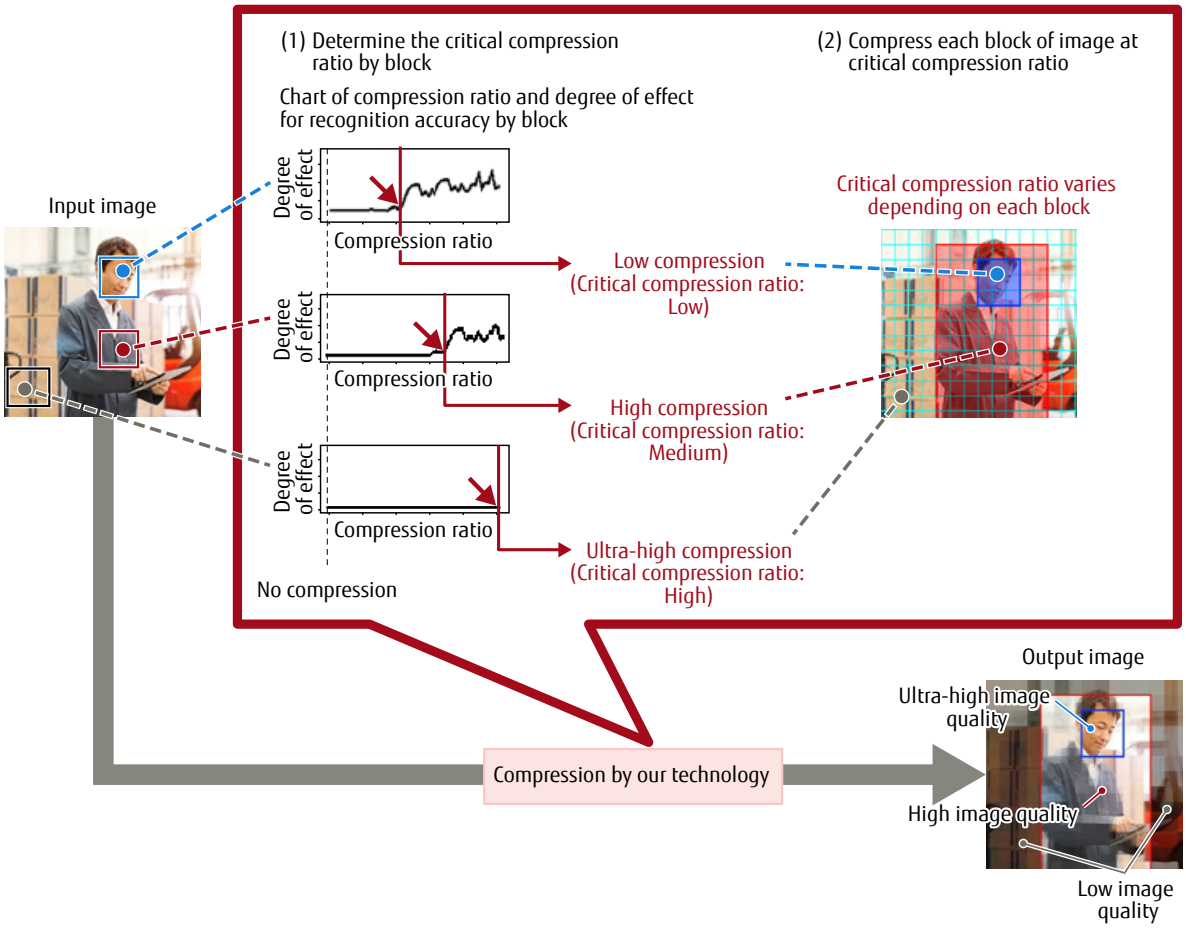


Figure 3
Determining critical compression ratio and compressing image.

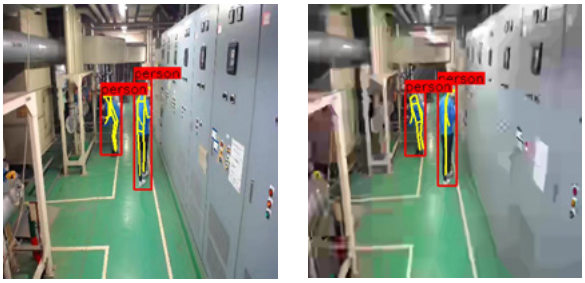
4.1. Evaluation of data size and recognition results

Two types of evaluation were performed based on the "Cascaded Pyramid Network (CPN)" [13] deep learning model, where the bone structures of people in the image cut out with YOLOv3 are recognized. First, a video image was compressed both at the compression ratio generally applied in conventional compression technology intended for visual confirmation by humans, and at the compression ratio determined using our new technology, and then the data sizes of the compressed images were compared. Next, the images decoded after compression were entered in the bone structure recognition process to compare the recognition results of both images.

4.2. Same recognition results achieved even with data size reduced to 1/7

The evaluation confirmed that the data size of the conventional compression technology intended for visual confirmation by humans was 11,814 bytes, whereas the data size using our technology was only 1,527 bytes, successfully reducing it to one seventh the data size.

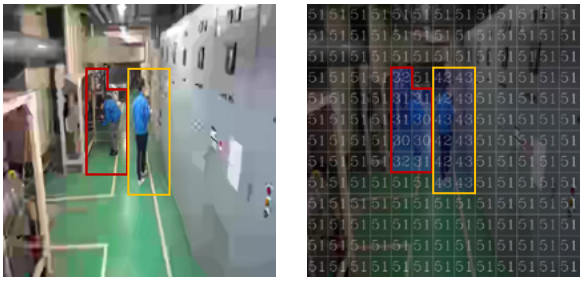
Figure 4 shows an example of an image where the bone structure information output through the bone structure recognition process is superimposed on each of the images decoded after conventional compression intended for visual confirmation by humans (a) and after compression using our technology (b). This evaluation confirmed that even when the data size using our technology was reduced to approximately one seventh the data size of the conventional compression technology intended for visual confirmation by



(a) Compression for visual confirmation by humans

(b) Compression using our technology

Figure 4
Results of bone structure recognition by compression for confirmation by humans and our technology.



(a) Image compressed by our technology

(b) Critical compression ratio by block

Note: The red and orange frames indicate zones that our technology has judged as important blocks for AI recognition.

Figure 5
Consideration of the compression process using our technology.

humans, the bone structures of the people could still be recognized in almost the same shapes as in the uncompressed video image in all video frames.

4.3. Considerations

This section describes key points for successfully reducing the data size to approximately one seventh, using the image of one frame extracted from a video (**Figure 5**).

Figure 5 (a) shows an image compressed using our technology and Figure 5 (b) shows the critical compression ratio for each block, which was calculated based on our technology. The compression ratio used was 0 to 51 within the range that can be set under H.265/HEVC. A larger value means a higher compression ratio.

Figure 5 (b) indicates that only the blocks that significantly affect bone structure recognition are set to low values ranging from 30 to 43, whereas all the others are set to 51, the highest value. The variations between 30 and 43 are based on results of the detailed assessment of the compression ratio by our technology, using the difference in the degree of effect for recognition accuracy, based on the method described in subsection 3.3. With the image converted at the conventional compression intended for visual confirmation by humans, on the other hand, the compression ratio is set to 30 for all of the blocks.

Therefore, only some areas (hereafter, zones) of each image are important for AI recognition. It can be said that the compression ratio can be increased more than before by setting unimportant zones to ultra-high compression, while changing the compression ratio for important zones even more finely.

However, it should be noted that the effects of compression may vary, depending on the image size, the number and scale of recognition targets, complexity of image movements, and other factors.

5. Summary and future plans

This article described the technology used to distinguish between important and unimportant zones of an image for AI recognition and how to control the compression ratio finely in order to optimize an image to a level of quality that can be recognized by AI and achieve high compression of video data.

This technology can be used to send a much smaller amount of video data than before from the field to the cloud, enabling AI analysis in the cloud environment. Even more advanced analysis and utilization can be achieved through combined use of analysis results with sensor data, document information and other information in the cloud, for example, by combining information about in-store customer behaviors based on video image analysis with sales information for marketing use.

In addition, the spread of the 5th-generation mobile communications system can provide advantages such as enabling easy relocation of installation locations of many cable-free cameras, which transmit video data wirelessly, when factory lines are changed, for instance. When combined with our technology, which produces the effect of high compression of video data captured by many cameras, 5G technology can lower

the threshold of use of ultra-high definition video images [14].

Fujitsu will continue to evaluate the new technology under various conditions assumed for practical applications, such as image capturing conditions for AI recognition targets as well as sizes and movements, so that the technology can be applied to various use cases, and further research and develop improved compression performance.

References and Notes

[1] Mobile communication system featuring high-speed, large-capacity, multiple simultaneous connections, and ultra-low delay.

[2] Ministry of Internal Affairs and Communications: White Paper 2018 —Information and Communications in Japan (2018).
https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2018/2018-index.html

[3] Center for Research and Development Strategy, Japan Science and Technology Agency: Panoramic View of the Systems and Information Science and Technology Field (2019) (in Japanese, Executive Summary in English).
https://www.jst.go.jp/crds/pdf/2018/FR/CRDS-FY2018-FR-02.pdf
https://www.jst.go.jp/crds/en/publications/CRDS-FY2018-FR-02.html

[4] Naohiko Hagiwara: Current Situation and Future Outlook of 5th-Generation Mobile Communication System (5G), Ministry of Internal Affairs and Communications (2019) (in Japanese).

[5] Ministry of Economy, Trade and Industry: 2017 Investigation Report on Infrastructure Development of Data-Driven Society in Japan (Analysis and Investigation Design of Factual Investigation of Information Processing) (2019) (in Japanese).
https://www.meti.go.jp/statistics/zyo/zyouhou/result-2/pdf/H29_report.pdf

[6] Fujitsu: Fujitsu Streamlines AI Video Recognition with High-Quality Compression Technology, (2020).
https://www.fujitsu.com/global/about/resources/news/press-releases/2020/0305-01.html

[7] FUJITSU JOURNAL: Video Data for AI Analysis Compressed to 1/10 the Data Size of Conventional Compression Technology –Cloud Applications to Contribute to Crime Prevention and Resolution of Other Social Issues, (2020) (in Japanese).
https://blog.global.fujitsu.com/jp/2020-06-04/01/

[8] ISO/IEC 23008-2:2017, High efficiency coding and media delivery in heterogeneous environments — Part2: High Efficiency Video Coding, Oct. 2017. | Recommendation ITU-T H.265 (2018), High Efficiency Video Coding (2018).

[9] Yoshihisa Yamada: Standard Moving Images for Video Encoding Evaluation, The Journal of the Institute of Image Information and Television engineers, Vol.62, No.8, pp. 1283–1285 (2008) (in Japanese).
https://www.jstage.jst.go.jp/article/itej/62/8/62_8_1283/_pdf/-char/ja

[10] J. Redmon, et al.: YOLOv3: An Incremental Improvement, CoRR, Vol. abs/1804.02767.
https://arxiv.org/pdf/1804.02767.pdf

[11] K. Simonyan, et al.: Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps, CoRR, Vol.1312.6034 (International Conference on Learning Representations (ICLR) 2014 Workshop) (2013).
https://arxiv.org/pdf/1312.6034.pdf

[12] David E. Rumelhart et al.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press (1986).

[13] Y. Chen, et al.: Cascaded Pyramid Network for Multi-Person Pose Estimation, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7103–7112 (2018).
https://arxiv.org/pdf/1711.07319.pdf

[14] Fujitsu: Local 5G, (2020) (in Japanese).
https://www.fujitsu.com/jp/innovation/5g/local5g/

Tomonori Kubota
Fujitsu Laboratories Ltd., Digital Innovation Core Unit
Mr. Kubota is currently engaged in research and development of real-time media processing system technology, AI-applied system technology, and Explainable AI (XAI) technology.

Takanori Nakao
Fujitsu Laboratories Ltd., Digital Innovation Core Unit
Mr. Nakao is currently engaged in research and development of video system technology, AI-applied system technology, and storage system technology.

Eiji Yoshida
Fujitsu Laboratories Ltd., ICT Systems Laboratory
Mr. Yoshida is currently engaged in research and development of data system technology.

Makoto Kubota
Fujitsu Laboratories Ltd., Digital Innovation Core Unit
Mr. Kubota is currently engaged in research and development of digital services using the edge cloud.