High-definition AI Inference Technology for Detecting a Wide Range of Objects with One Camera at One Time

Hiroyuki Uzawa, Shuhei Yoshida, Yuukou Iinuma, Saki Hatta, Daisuke Kobayashi, Yuya Omori, Yusuke Horishita, Tomoki Onoe, Ken Nakamura, and Koji Yamazaki

Abstract

Object detection in high-definition video is required in video artificial intelligence (AI) applications for edge/terminals to detect a wide range of objects with one camera at one time. Although various AI inference schemes for object detection (e.g., You Only Look Once (YOLO)) have been proposed, they have a limitation regarding the input image size, thus need to shrink the input high-definition image into that limited size. This collapses small objects, making them undetectable. This article introduces high-definition AI inference technology we previously proposed for solving this problem, with which multiple object detectors cooperate to detect small and large objects in high-definition video.

Keywords: AI, object detection, high-definition video

1. Introduction

Object detection for identifying and locating objects in images plays an important role in video artificial intelligence (AI) applications for edge/terminals. One application is automated beyond-visual-line-of-sight (BVLOS) drone flight. BVLOS flight means flying beyond the operator's visual range and enables a drone to cover far greater distances. For BVLOS flight, a drone has to detect objects in images input from a mounted camera in real time for safe flight, especially to avoid flying directly over passersby or cars under the flight path. Another application is camera surveillance, and the object detection for determining suspicious people from a crowd of people has to be done in a terminal to comply with personal-information-protection requirements, such as the General Data Protection Regulation (GDPR). Another application is road-traffic monitoring.

Object detection in high-definition video enables detecting a wide range of objects with a single camera at one time. This makes it possible to detect passersby and cars under the flight path from higher altitude for automated BVLOS drone flight and detect suspicious people from a greater distance for camera surveillance. In a high-definition image such as full HD and 4K, since objects near and far from the mounted camera can coexist in the same image due to a wide angle of view, both large and small objects can be included in the image. Therefore, object detection in high-definition video has to be able to detect not only large but also small objects with high accuracy.

Various AI inference schemes for object detection have been proposed, such as You Only Look Once

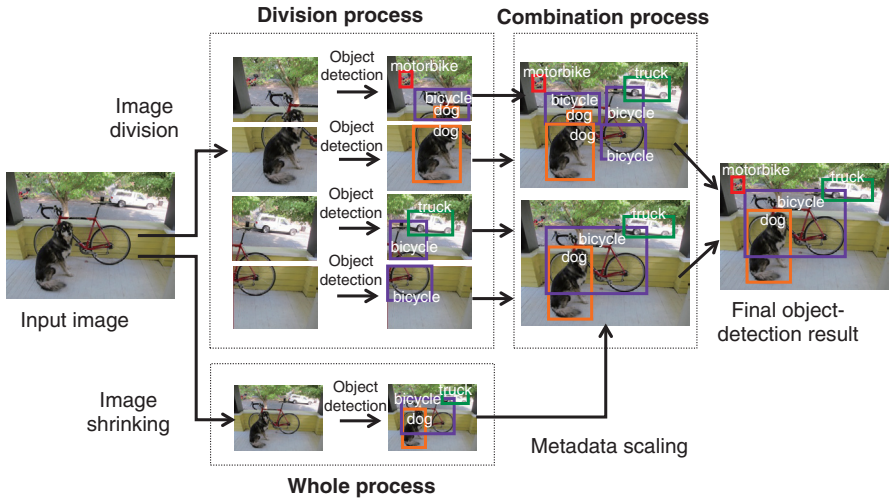


Fig. 1. Our high-definition AI inference technology for object detection.

(YOLO) [1] and Single Shot MultiBox Detector (SSD) [2]. Such an inference scheme has a convolutional neural network (CNN) that predicts object bounding boxes and class probabilities in a single evaluation. The input-layer size of a trainable CNN is limited to suppress the computational complexity and training-process difficulty. For example, the largest input-layer size in the standard YOLO model YOLOv3 is 608 × 608 pixels. Thus, even if a full HD image is input, the input image is shrunk to the limited input-layer size and inference is executed with the shrunk input image. Large objects in the input image can be detected, but small objects are collapsed and difficult to detect. Dividing the input image into a limited size can also be considered to prevent shrinking the input image [3–8], but this means large objects that straddle the divided images cannot be detected because the characteristic parts for identifying objects are also divided. In other words, such schemes are unsuitable for object detection in high-definition video.

We introduce high-definition AI inference technology we previously proposed [9, 10] for solving this problem, with which multiple object detectors cooperate to detect small and large objects in high-definition images. This technology is suitable for hardware implementation because all object detectors can be executed in parallel for real-time operation. In addition, any AI inference scheme for object detection can be applied, and re-training for applying our technology is not necessary.

The remainder of the article is organized as follows. In Section 2, we give details of our technology and present the results of an evaluation of the technology in Section 3. We conclude this article in Section 4.

2. Our high-definition AI inference technology

2.1 Overview of our technology

Our high-definition AI inference technology [9, 10] enables both small and large objects to be detected in high-definition images through cooperation among multiple object detectors (**Fig. 1**). An input image is divided into images of limited size, and objects are detected on the basis of an inference scheme for every divided image. In parallel, object detection is also done in the whole image shrunk to a limited size to detect large objects that straddle the divided images. By combining these object-detection results, a final object-detection result is obtained. In other words, the undetected objects in the shrunk whole image are interpolated from the object-detection results obtained from the divided images. Therefore, both small and large objects can be detected in high-definition images.

We explain the mechanism by which high-definition object detection can be achieved with our technology by using YOLO as an example scheme. YOLO has a CNN for detecting objects included in an input image. The CNN divides the input image into $S \times S$ grids and predicts bounding boxes and class probabilities for each grid. The resolution for detecting objects depends on only the grid size. Our technology reduces the grid size. **Figure 2** shows an

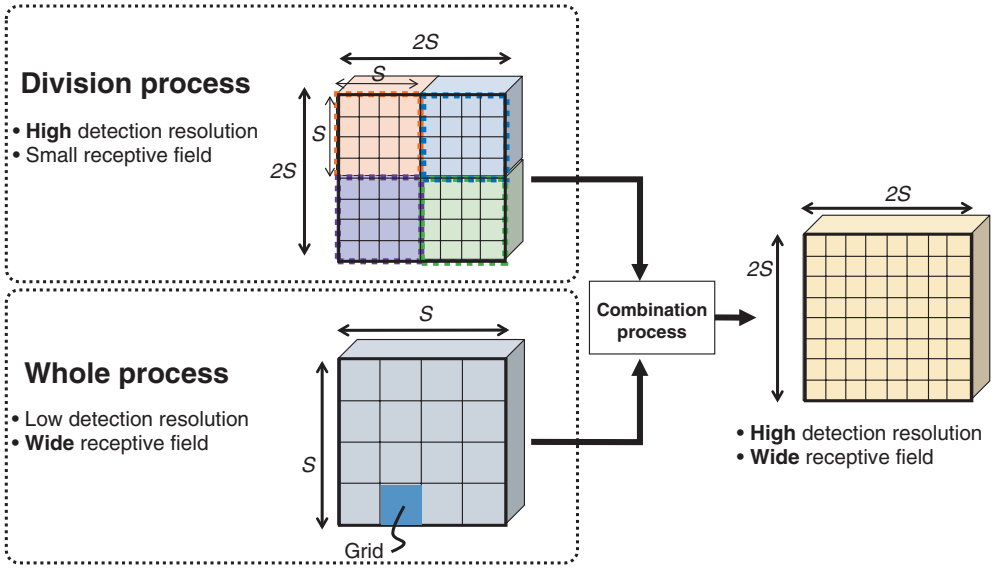


Fig. 2. High-definition object-detection mechanism.

example in which the input image is divided into four images. In the division process, since each divided image has $S \times S$ grids, the four total divided images can be regarded as $2S \times 2S$ grids. This means that the grid size is reduced to half. However, the division process causes the receptive field to narrow because it does not execute the convolutional operations for the whole image, thus cannot detect large objects. In contrast, in the whole process, although the detection resolution is low, a wide receptive field can be obtained. Therefore, with our technology, the division process with high detection resolution and small receptive field and the whole process with low detection resolution and wide receptive field are combined, which enables object detection with high detection resolution and wide receptive field. Concretely, when an image with 1920 × 1080 pixels is input, the minimum gird size without our technology (YOLOv3 only) is 60 × 34 pixels because S is 32. In contrast, with our technology, the minimum grid size becomes 30 × 17 pixels when there are four divisions, and higher detection resolution can be provided.

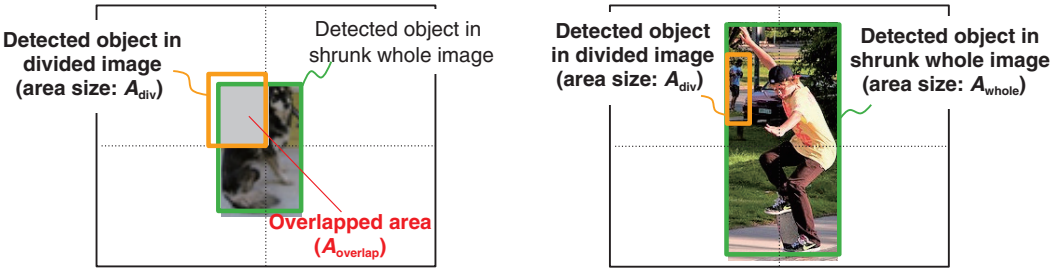
Our technology is suitable for hardware implementation because all object-detection processes can be executed in parallel. The same weight coefficients can also be used among all object-detection processes. In addition, any AI inference scheme for object detection can be applied, and re-training for applying our technology is not necessary. Moreover, the computational complexity with our technology is proportional to the number of divisions. This means that our technology can reduce the grid size with less complexity than increasing the number of grids in the CNN because the complexity of the CNN increases with the square of the number of grids.

2.2 Details of combination process

In the combination process, the undetected objects in the shrunk whole image are selected from the object-detection results obtained with the divided images. The detected objects in the shrunk whole image and the selected objects are output as the final object-detection results.

This selection requires determining whether an object detected in the divided image is an undetected one in the shrunk whole image. The combination process calculates two indicators: a multiplicity between the detected object in the divided image and that in the shrunk whole image and the ratio between the area size of the detected object in the divided image and that in the shrunk whole image.

When both indicators are high, the process determines that the detected object in the divided image is the same as that in the shrunk whole image and excludes it from the objects to be selected. In contrast, when either of these indicators is low, the process determines that the detected object in the divided image is different from that in the shrunk whole image. This is executed for all detected objects in the divided images in combination with all detected



(a) $A_{overlap}$ and A_{div} (b) A_{div} and A_{whole}

Fig. 3. Examples of A_{div}, A_{whole}, and $A_{overlap}$.

objects in the shrunk whole image, and the non-excluded detected objects in the divided images are selected as the undetected objects in the shrunk whole image. Moreover, the object with the class detected only in the divided images is selected when its object size is sufficiently small that it cannot be detected in the whole shrunk image, e.g., its object size is less than 1/100 of the input image size.

The above-mentioned multiplicity $I_{multiplicity}$ is a dedicated indicator for enabling comparison between the divided and whole objects and is expressed as

$$I_{multiplicity} = A_{overlap} / A_{div}, \quad (1)$$

where $A_{overlap}$ is the overlapped area size between the divided object detected in the divided image and the whole object detected in the shrunk whole image, and A_{div} is the area size of the divided object, as shown in **Fig. 3(a)**. Although intersection over union (IoU) is often used as multiplicity, it is unsuitable to use to compare a divided object and a whole object because it assumes comparison between whole objects. Therefore, we newly define $I_{multiplicity}$. When $I_{multiplicity}$ is larger than or equal to a threshold α, the multiplicity is determined to be high. Under the high multiplicity condition, the divided object is likely to be the same as the whole object.

Even if the multiplicity is high, the detected object in the divided image may be different from that in the shrunk whole image when the area-size ratio between those objects is high, as shown in **Fig. 3(b)**. The above-mentioned size ratio I_{ratio} is used for determining this condition. The I_{ratio} is given by

$$I_{ratio} = A_{div} / A_{whole}, \quad (2)$$

where A_{whole} is the area size of the detected object in the shrunk whole image. Under the high multiplicity condition, when I_{ratio} is larger than or equal to a threshold β, the detected object in the divided image is determined to be same as that in the shrunk whole image.

With these indicators in the same class, the undetected objects in the shrunk whole image are selected from the detected objects in the divided images, and both large and small objects can be detected while avoiding duplicate detection of the same object.

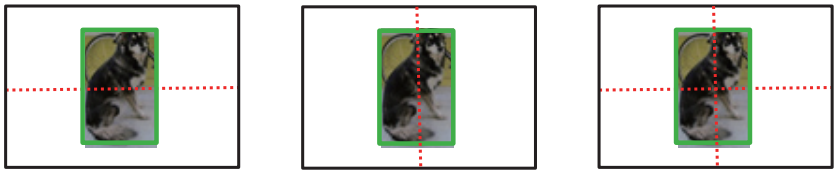
3. Object-detection performance

We applied our technology to the standard YOLOv3 model with 608 × 608 pixels and evaluated object-detection performance. The same weight coefficients are used between all object detectors in our technology, as described in Section 2.1. We used the weight coefficients published in [11]. These coefficients are pre-trained with the Microsoft Common Objects in Context (MS COCO) dataset [12], which is a widely used object-detection dataset with 80 classes.

3.1 Optimization of pre-set parameters

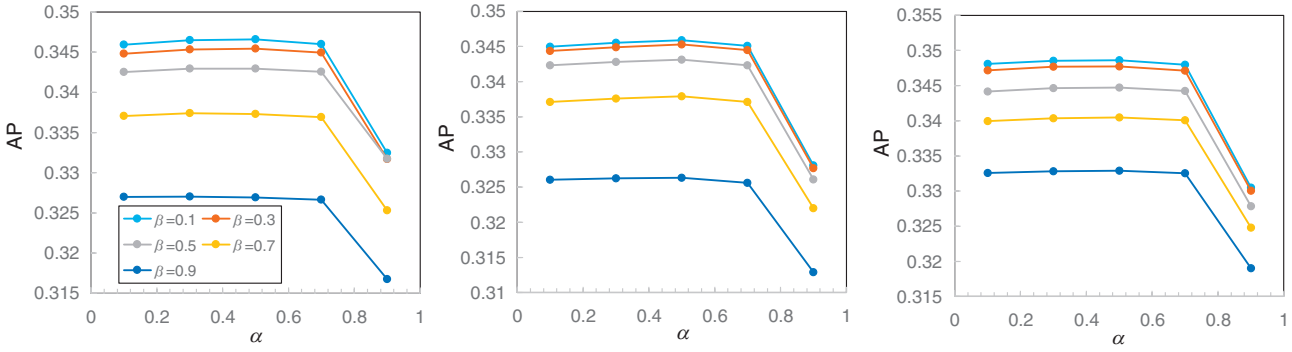
With our technology, α and β are pre-set for the combination process. An object detected in the divided image is more likely to be determined as undetected in the shrunk whole image when these thresholds are higher and more likely to be determined as the same when these thresholds are lower. Thus, α and β should be optimized to execute the combination process properly.

Object-division patterns can be mainly classified into horizontal division (**Fig. 4(a)**), vertical division (**Fig. 4(b)**), or cross-shaped division (**Fig. 4(c)**). When there are two divisions, only the vertical or horizontal division pattern can occur. In contrast, when there are more than two divisions, all division patterns can occur, and the division pattern for the



(a) Horizontal division (b) Vertical division (c) Cross-shaped division

Fig. 4. Division patterns.



(a) Horizontal division (b) Vertical division (c) Cross-shaped division

Fig. 5. Relationship between AP, α, and β.

object varies in accordance with the position of that object in the image. For this reason, commonly available α and β need to be obtained for various division patterns.

We searched for the optimized α and β where the average precision (AP) is maximized. The AP is obtained from precision and recall. In this search, α and β were varied from 0.1 to 0.9 in increments of 0.2. We used the images in the MS COCO dataset val2017 because the objects included in these images are large and the division patterns shown in Fig. 4 can be reliably generated.

Figure 5 shows the measured relationship between AP, α, and β. The AP decreased as β increased from 0.1 in all division patterns. The AP also reached its maximum when α was 0.5 in all division patterns. The AP gradually decreased as α increased from 0.5. This is because an object detected in the divided image is more likely to be determined as undetected in the shrunk whole image as α becomes higher. In other words, the optimized α and β are 0.5 and 0.1, respectively, for all division patterns. Therefore, the combination process with our technology can be properly executed by pre-setting α to 0.5 and β to 0.1.

3.2 Object-detection performance

We conducted evaluations to determine the effectiveness of our technology. On the basis of the optimized result described in the previous section, α and β were set to 0.5 and 0.1, respectively. There are two divisions.

We first conducted a basic evaluation using 5000 images in the MS COCO dataset val2017 and measured the AP for each class. Although the images included in the MS COCO dataset are standard definition (SD) images, such as 600 × 400 pixels, large objects account for a higher percentage of the objects in the images; thus, we can determine the image-division penalty with our technology by comparing the APs with and without our technology. The measured AP is that averaged by 10 IoU thresholds in 0.05 increments from 0.5 to 0.95.

Figure 6 shows the measurement results. The AP improved in almost all classes and by a maximum of 1.2 times. This means that our technology can suppress the image-division penalty for large objects and

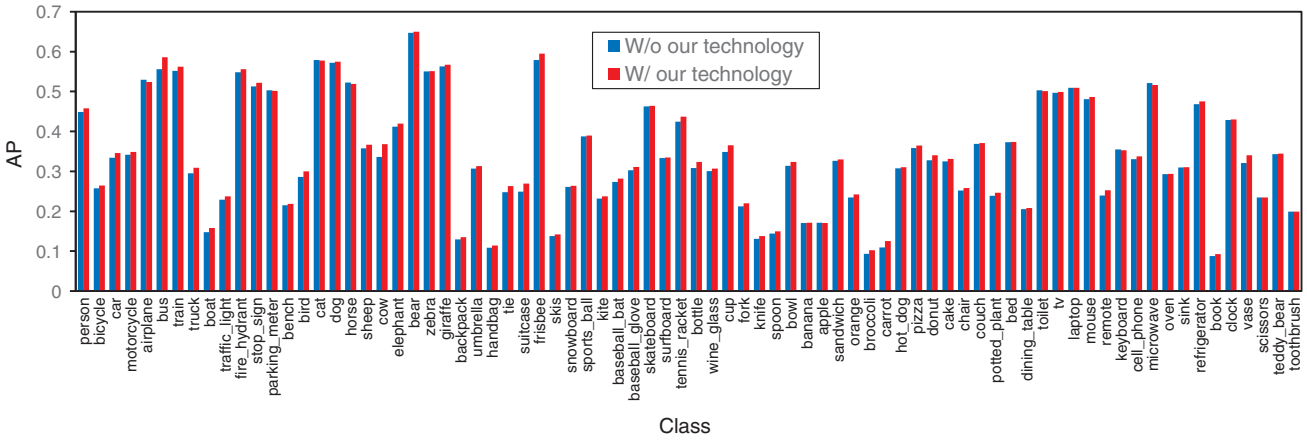


Fig. 6. AP of each class (image size: about 600 × 400 pixels).

Table 1. Summary of the detection results (image size: 1920 × 1080 pixels).

Class	W/o our technology (standard YOLOv3 model only)		W/ our technology	
	Precision [%]	Total number of objects detected in 150 images	Precision [%]	Total number of objects detected in 150 images
Person	62.5	176	62.8	298
Car	83.9	1426	81.5	2049
Bicycle	50.0	6	54.5	11
Truck	17.1	111	19.3	176
Bus	33.3	6	15.0	20
Motorbike	61.5	13	59.5	37

improve object-detection performance even in SD images.

We then evaluated object-detection performance with full HD images. We selected 150 full HD (1920 × 1080) images from the VisDrone2019 dataset [13] and evaluated precision and the number of detected objects. The VisDrone dataset is a large-scale drone-captured dataset with ten classes and includes full HD images. For the weight coefficients, we used the same coefficients pre-trained using the MS COCO dataset as above. To calculate precision, we remapped people and pedestrian labels to person label and van label to car label in the VisDrone dataset, and only the common class labels between MS COCO and VisDrone were evaluated.

Table 1 summarizes the detection results, and **Figure 7** shows the example images obtained in this evaluation. From Table 1, our technology enabled the number of detected objects to be increased while maintaining precision. For example, the number of detected objects in the person class was 1.7 times higher with our technology than without it. Across all evaluation classes, it was 2.1 times higher on average. As shown in Fig. 7, small objects such as passersby and distant cars could be detected with our technology but could not be detected without it.

Figure 8 shows the size distribution of detected objects. Our technology could detect much smaller objects. Specifically, the minimum width size was halved from 12 pixels without our technology to 6 pixels with it. This is because $2S$ (width) × S (height) grids are achieved with our technology when there are two divided images in the division process, as described in Section 2.

These results indicate that our technology can improve object-detection performance in not only SD images but also HD images by suppressing the image-division penalty.

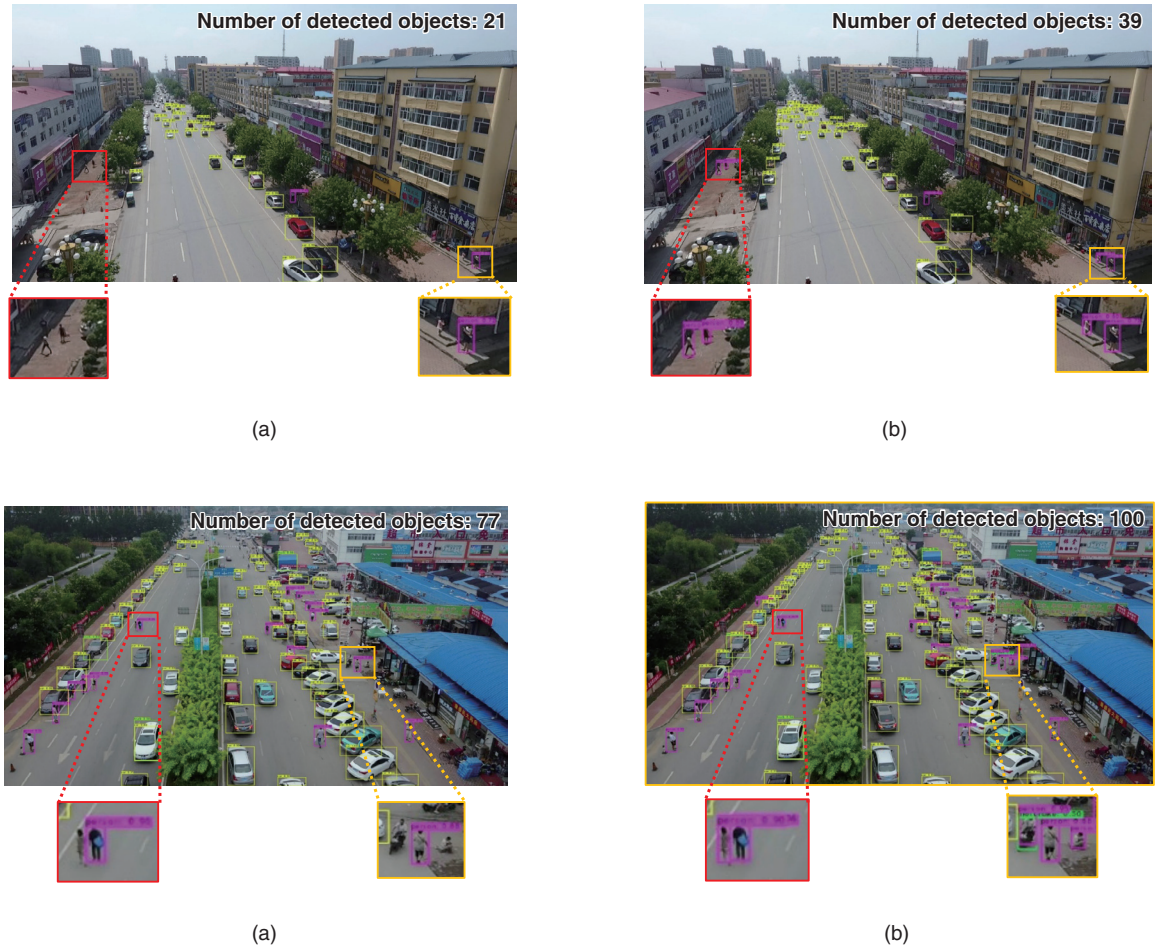


Fig. 7. Examples of detection results: (a) without our technology (standard YOLOv3 model only) and (b) with our technology.

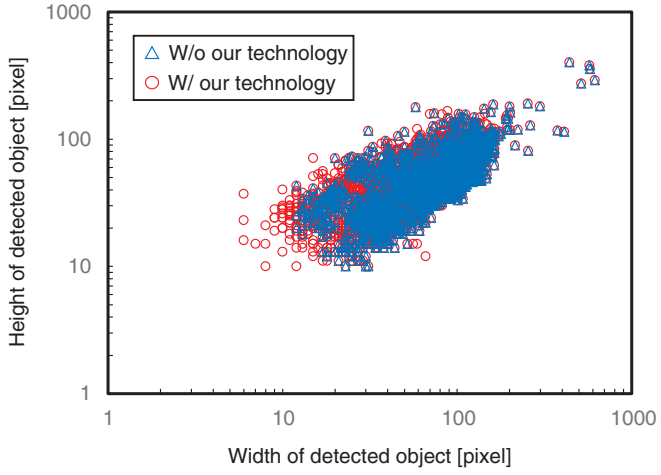


Fig. 8. Size distribution of detected objects in high-definition images.

4. Conclusion

This article introduced our technology that enables the detecting of a wide range of objects with one high-definition camera at one time for edge/terminal AI applications. It detects objects with a high detection resolution and wide reception field by combining the division process with multiple divided images and the whole process with the shrunk whole image.

We applied our technology to the standard YOLOv3 model with 608 × 608 pixels and evaluated object-detection performance. The evaluation results indicate that our technology can improve object-detection performance in not only standard definition images but also high-definition images while suppressing the image-penalty.

This technology is suitable for hardware implementation because all object detectors can be executed in parallel for real-time operation. Any AI inference scheme for object detection can be applied, and retraining for applying our technology is not necessary. This will facilitate its application to various edge/terminal AI applications.

References

[1] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv:1804.02767, 2018. https://doi.org/10.48550/arXiv.1804.02767

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," arXiv:1512.02325, 2015. https://doi.org/10.48550/arXiv.1512.02325

[3] V. Růžička and F. Franchetti, "Fast and Accurate Object Detection in High Resolution 4K and 8K Video Using GPUs," Proc. of 22nd Annual IEEE High Performance Extreme Computing Conference (HPEC 2018), Waltham, MA, USA, Sept. 2018. https://doi.org/10.1109/HPEC.2018.8547574

[4] G. Plastiras, C. Kyrkou, and T. Theocharides, "Efficient ConvNet-based Object Detection for Unmanned Aerial Vehicles by Selective Tile Processing," Proc. of the 12th International Conference on Distributed Smart Cameras (ICDSC 2018), Eindhoven, The Netherlands, Sept. 2018. https://doi.org/10.1145/3243394.3243692

[5] D. Vorobjov, I. Zakharava, R. Bohush, and S. Ablameyko, "An Effective Object Detection Algorithm for High Resolution Video by Using Convolutional Neural Network," Advances in Neural Networks, LNCS, Vol. 10878, pp. 503–510, 2018. https://doi.org/10.1007/978-3-319-92537-0_58

[6] R. Bohush, S. Ablameyko, S. Ihnatsyeva, and Y. Adamovskiy, "Object Detection Algorithm for High Resolution Images Based on Convolutional Neural Network and Multiscale Processing," Proc. of the 4th International Workshop on Computer Modeling and Intelligent Systems (CMIS-2021), Zaporizhzhia, Ukraine, Apr. 2021. https://doi.org/10.32782/cmis/2864-12

[7] M. Gao, R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Dynamic Zoom-in Network for Fast Object Detection in Large Images," Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, June 2018, pp. 6926–6935. https://doi.org/10.1109/CVPR.2018.00724

[8] C. Tang, Y. Ling, X. Yang, and W. Jin, "Multi-view Object Detection Based on Deep Learning," Applied Sciences, Vol. 8, No. 9, p. 1423, 2018. https://doi.org/10.3390/app8091423

[9] H. Uzawa, S. Yoshida, Y. Iinuma, S. Hatta, D. Kobayashi, Y. Omori, K. Nakamura, S. Takada, H. Toorabally, and K Sano, "High-definition Object Detection Technology Based on AI Inference Scheme and Its Implementation," IEICE Electronics Express, Vol. 18, No. 22, P. 20210323, 2021. https://doi.org/10.1587/elex.18.20210323

[10] H. Uzawa, S. Yoshida, Y. Iinuma, S. Hatta, D. Kobayashi, Y. Omori, Y. Horishita, K. Nakamura, S. Takada, H. Toorabally, K. Nitta, K. Yamazaki, and K. Sano, "High-definition Technology of AI Inference Scheme for Object Detection on Edge/Terminal," IEICE Electronics Express, Vol. 20, No. 13, p. 20232002, 2023. https://doi.org/10.1587/elex.20.20232002

[11] Website of YOLO, https://pjreddie.com/darknet/yolo/

[12] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," Computer Vision – ECCV 2014, 2014, LNCS, Vol. 8693, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[13] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision Meets Drones: A Challenge," arXiv:1804.07437, Apr. 2018. https://doi.org/10.48550/arXiv.1804.07437

Hiroyuki Uzawa

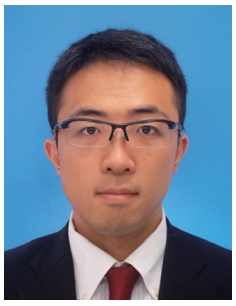
Senior Manager, NTT Device Innovation Center.

He received a B.E. and M.E. in electrical engineering from Tokyo University of Science in 2006 and 2008. In 2008, he joined NTT Microsystem Integration Laboratories, where he was engaged in research and development of the design techniques for network system on a chip (SoC). Since 2019, he has been engaged in research and development of a high-definition AI inference engine at NTT Device Innovation Center. He received the 2012 IEICE Young Engineer Award and the 2021 IEICE ELEX Best Paper Award. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Shuhei Yoshida

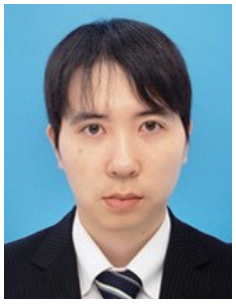
Research Engineer, NTT Device Innovation Center.

He received a B.E. and M.E. in computer science and systems engineering from Kobe University, Hyogo, in 2014 and 2016. In 2016 he joined NTT Device Innovation Center and has been engaged in research and development on hardware design methodology, field-programmable gate array (FPGA)-based network-traffic-monitoring systems, and a high-definition AI inference engine. He is a member of IEICE.



Yuukou Iinuma
Engineer, NTT Device Innovation Center.
He received a B.E., and M.E. in information science and technology from the University of Tokyo in 2019 and 2021. He joined NTT in 2021 and is currently engaged in research and development of an AI inference engine for high-definition video and object-detection algorithms in 8K video.

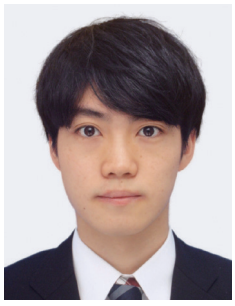
Saki Hatta
Senior Manager, NTT Device Innovation Center.
She received a B.S. and M.S. in material engineering from Tokyo Institute of Technology in 2009 and 2011. In 2011, she joined NTT Microsystem Integration Laboratories, where she was engaged in the research and development of the design techniques for a network SoC. She is currently with NTT Device Innovation Center and engaged in research and development of a network-traffic-monitoring system. She was a recipient of the IEEE CEDA 2018 Young Researcher Award. She is a member of IEICE.



Daisuke Kobayashi
Senior Manager, NTT Device Innovation Center.
He received a B.E. and M.E. in information and communication engineering from the University of Electro-Communications, Tokyo, in 2007 and 2009. He joined NTT Cyber Space Laboratories in 2009 and since then has been engaged in research and development on high quality video coding and transmission. Since 2020, he has been engaged in research and development of an AI inference hardware engine.

Yuya Omori
Research Engineer, NTT Device Innovation Center.
He received a B.E. and M.E. from the University of Tokyo in 2012 and 2014. In 2014, he joined NTT Media Intelligence Laboratories and has been engaged in research and development on parallel processing architecture and a high efficiency algorithm of video-processing hardware.

Yusuke Horishita
Senior Research Engineer, NTT Device Innovation Center.
He received a B.E. and M.E. in science and technology from Keio University, Tokyo, in 2010 and 2012. In 2012, he joined Canon Inc., where he was engaged in the development of SoCs for digital cameras and multi-function printers. Since 2022 he has been involved in research and development of an AI processor at NTT Device Innovation Center. He is a member of IEICE.



Tomoki Onoe
Researcher, NTT Device Innovation Center.
He received a B.S. and M.S. from Kyushu University, Fukuoka, in 2021 and 2023. In 2023, he joined NTT Device Innovation Center and has been engaged in research and development on video AI inference large-scale integration (LSI).

Ken Nakamura
Director, NTT Device Innovation Center.
He received a B.E. and M.E. in instrumentation engineering from Keio University, Kanagawa, in 1995 and 1997, and received a Ph.D. in information sciences from Tohoku University, Miyagi, in 2023. In 1997, he joined NTT Human Interface Laboratories and has been engaged in research and development on video coding LSI and systems. He is a member of IEICE and the Information Processing Society of Japan (IPSJ).

Koji Yamazaki
Director, NTT Device Innovation Center.
He received a B.S. in environmental information science from Keio University, Kanagawa, in 2000 and received a Master of Arts and Sciences (information studies) in interdisciplinary information science from the University of Tokyo in 2004. In 1998, he founded a venture company in Tokyo, Japan. In 2004, he joined NTT Microsystem Integration Laboratories. He has been engaged in research and development of LSIs and related FPGA accelerated systems.