Real-time Extraction of Objects from Any Background Using Machine Learning

Hirokazu Kakinuma, Jiro Nagao, Hiromu Miyashita, Yoshihide Tonomura, Hidenobu Nagata, and Kota Hidaka

Abstract

NTT Service Evolution Laboratories is conducting research and development on the Kirari! ultra-realistic communication system, which can make an athlete or performer in a remote location seem to be right before the viewer's eyes. This article describes a system that can extract objects from an arbitrary background in real time, which is essential for realistic remote presentation of athletes and performers using pseudo-three-dimensional video and other techniques.

Keywords: image segmentation, machine learning, ultra-realism

1. Introduction

The ability to accurately identify the region of a person or object within an image is an essential technology for performing high-quality image editing and composition, and it is therefore a major research theme in computer vision. Selecting the object region is also essential for achieving realism in the Kirari! ultra-realistic communication system when performing pseudo-three-dimensional video display. In the article "Real-time Extraction of Objects with Arbitrary Backgrounds" [1], NTT Service Evolution Laboratories proposed a system able to extract only the object region in real time from video of a sports venue or performance stage, without using studio equipment such as a green screen. This article introduces a system that can extract object regions with greater accuracy. It was developed by introducing machine learning to distinguish more subtle differences in feature values that have been indistinguishable earlier and generating feature values of the object being extracted using infrared light.

2. Framework for real-time object extraction using machine learning

Background subtraction is a common method for extracting objects in real time. Background subtraction involves finding the differences between the input image and a background image and applying a threshold to identify changes as the object area. This method is fast and requires little preparation, so it is widely used. However, there are challenges with this method, including the difficulty in deciding an appropriate threshold value and the inability to support backgrounds that change.

We have developed an object extraction method that uses a neural network (NN) to convert input feature vectors to a different feature space and perform the discrimination. With the NN, we hope to derive a feature space for making discriminations within the NN using training data provided beforehand and to automatically convert to this more suitable feature space. Inputs other than the image being detected, for example, reference images with features of the object, images from different times, region information, and infrared images, can also be included and will also be converted to suitable feature spaces in the

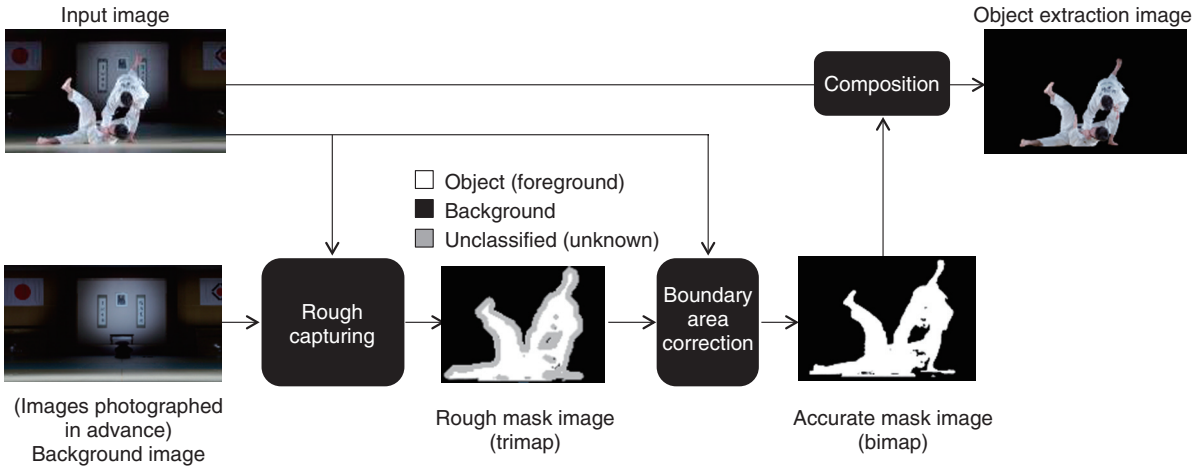
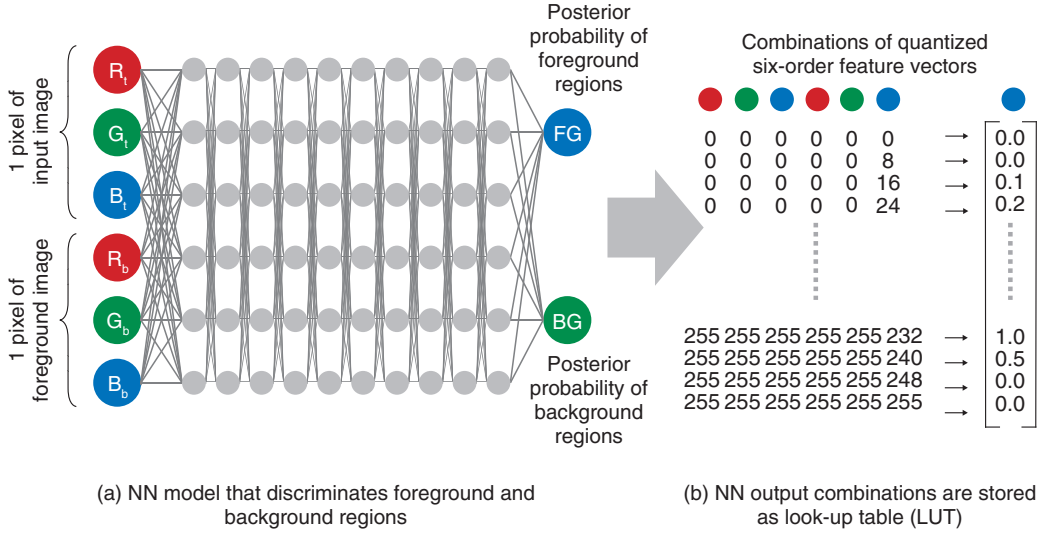


Fig. 1. Framework for real-time object extraction using machine learning.



(a) NN model that discriminates foreground and background regions

(b) NN output combinations are stored as look-up table (LUT)

Fig. 2. Training process.

NN in order to perform object extraction using higher-order feature values of the background and object. This should make operations such as changing the background more robust.

The workflow for the system we developed is shown in **Fig. 1**. Object extraction is done in two steps. In the first step, the object region is selected using a rough mask image (a trimap[*1]), and in the second step, a matting process[*2] dependent on this trimap creates a more accurate object region. Machine learning is used to capture trimap in the first step.

Machine learning is divided broadly into training and application processes. In the training process, parameters in the NN model are learned from training data. This process is shown in **Fig. 2**. Training data are first prepared. Background images that do not contain the object, and sample images that do contain

*1 Trimap: A region map indicating known and unknown regions of an image. Known foreground regions are set to white, known background regions to black, and unknown regions to grey.

*2 Matting process: A process that derives an alpha mask for extracting the object. The alpha mask has values ranging from 0 to 1, and the extracted image is obtained by multiplying the input image by the mask.

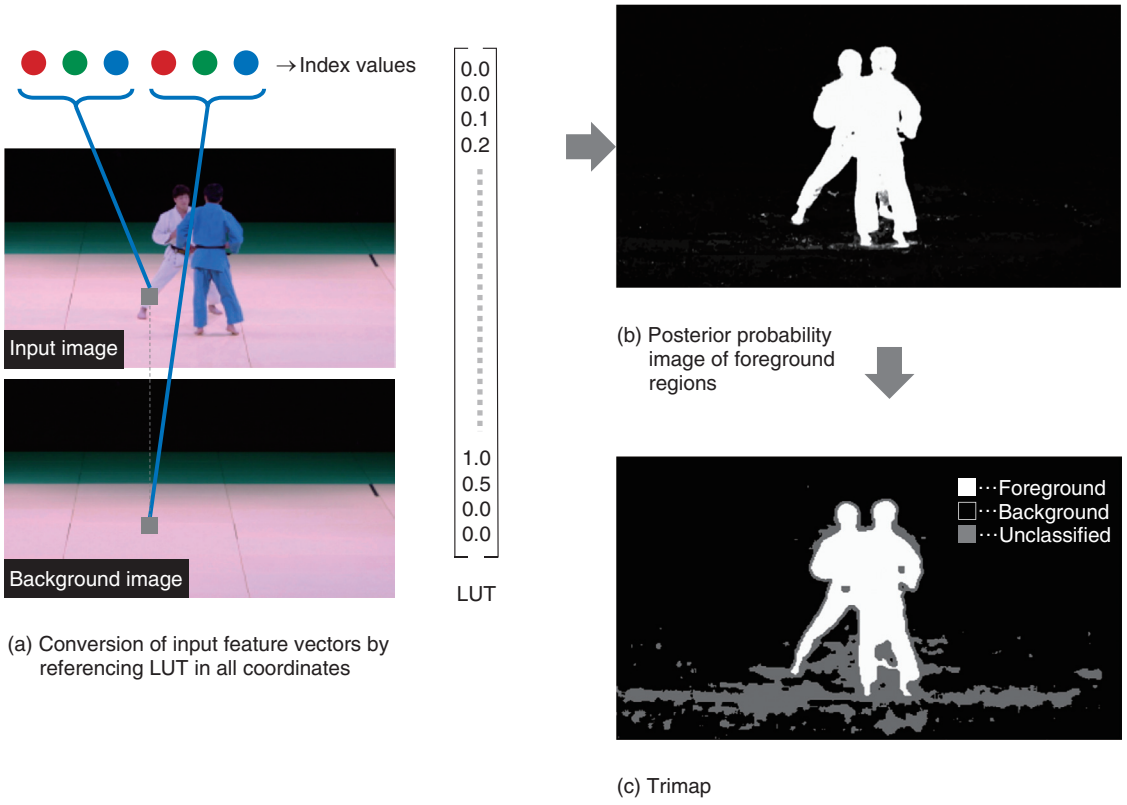


Fig. 3. Application process.

the object are obtained, and correct mask images are created. Then, input feature vectors are created. These feature vectors combine target pixels from the sample image corresponding to the foreground area in the mask, with the corresponding pixels from the background image. These combinations are used to train for the foreground region. Similarly, input feature vectors combining the background pixels from the sample image and corresponding pixels from the background image are created, and these combinations are used to train for the background region.

In this way, we obtained an NN model able to discriminate the foreground from the background regions for combinations of input target pixels and background image pixels (Fig. 2(a)). Generally, NN processing requires large computing resources, so we increased speed by implementing processing using a look-up table (LUT). We reduced the number of gradations in the input feature vectors by quantizing them and stored all combinations of quantized input feature vectors and NN outputs as an LUT (Fig. 2(b)). Note that in Fig. 2, we describe this process in terms of RGB (red, green, and blue color model) pixels for simplicity, but the input vectors can include more than color information, for example, image position.

The process for applying the LUT to generate the trimap is shown in **Fig. 3**. A quantized input feature vector is derived in a process similar to that for the machine learning process, and this derived feature vector is used to reference the LUT, rapidly determining a posterior probability that the pixel in question is in the foreground (Fig. 3(b)). The trimap is generated from the derived foreground-posterior-probability image by setting regions that are not clearly foreground or background to the unclassified region (Fig. 3(c)). The unclassified region in the trimap is discriminated by using a nearest-neighbor search with information about whether pixels neighboring the pixel in question and having similar feature vectors were classified as foreground or background.

The details of this nearest-neighbor search are shown in **Fig. 4**. For each pixel in the unclassified region, a spiral search in the local neighborhood is done to determine whether the pixel is more similar to a foreground pixel or a background pixel, and this is used to derive an alpha value for the pixel. The

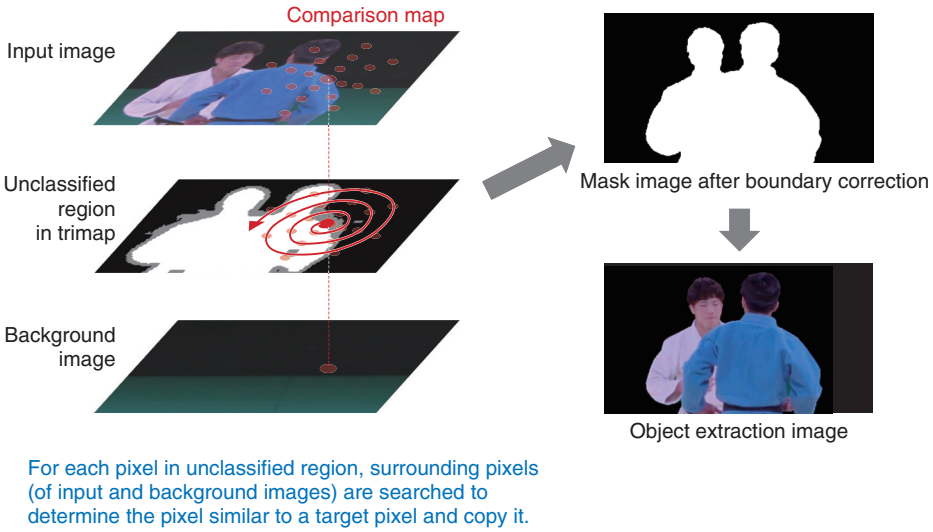


Fig. 4. Boundary correction by nearest-neighbor search, and masking process.

object is then extracted using the derived alpha values. By introducing this boundary correction process, we can extract objects based on information from the pixel itself and also from surrounding pixels. We also provide a framework for optimizing this process such as by processing the rough mask on a low-resolution image, or by performing only the magnification process on all pixels.

3. Real-time extraction of objects from backgrounds of similar color using infrared

Even though machine learning is used to automatically convert input feature vectors to a high-order feature space within the NN, we are still extracting objects based on color and shape information, so it is theoretically impossible to separate cases when the input feature vectors are the same. For this reason, we developed an object extraction system using RGB and infrared (IR) cameras, utilizing IR light that is invisible to the naked eye in an attempt to add new features.

The object extraction photography environment using the RGB and IR cameras is shown in **Fig. 5**. An RGB camera and an IR camera were placed side by side, respectively capturing visible light and IR images. The background was illuminated with IR, while deliberately preventing IR light from reflecting from the object. In this way, the background appeared brighter in the IR image, while the object was relatively dark.

In the RGB camera image in Fig. 5, the background and the person are of the same color, so it is difficult to distinguish them based on color, but the person's silhouette can be obtained from the IR camera image. Then, to extract the object accurately from the IR camera image, we correct for the parallax between the IR camera and the RGB camera. To perform the correction, photos of a calibration board were taken beforehand, a projection transform from the IR image was derived so that the same feature points from the IR camera and the RGB camera aligned, and this transform was applied. By adding the IR image obtained in this way to the input of machine learning, we were able to extract objects using new features other than color and shape.

4. Evaluation experiments

We developed a real-time object extraction system using machine learning and used it in the Cho-Kabuki play called "Tsumoru Omoi Hana no Kaomise," at the Niconico Chokaigi event held in April 2018 at Makuhari Messe, Chiba, Japan. The system developed is capable of 3840 × 2160 resolution video at a frame rate of 60 fps, but for this trial, a resolution of 1920 × 1080 and a frame rate of 59.94 fps were used for compatibility with other systems. In the climax scene of the Cho-Kabuki main production, Koretaka-shinno, played by Shido Nakamura II, emerges from the screen in real time as the stage video background changes, while the confrontation with Princess

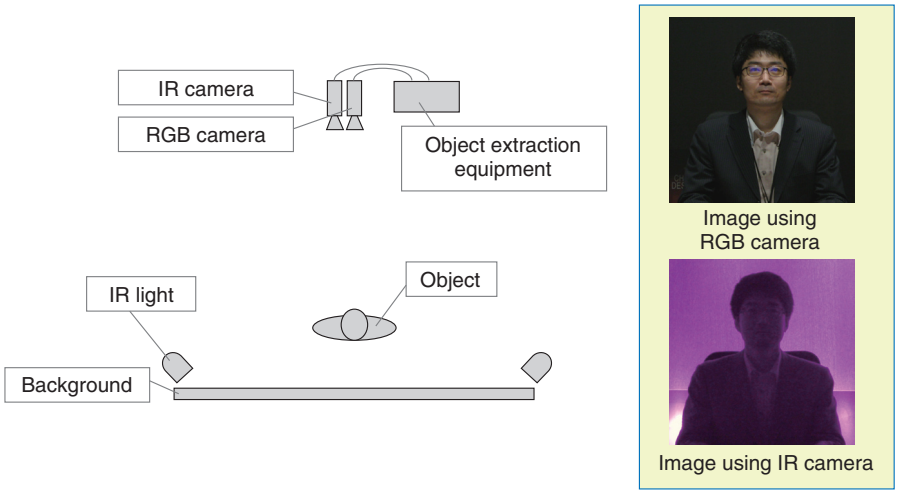
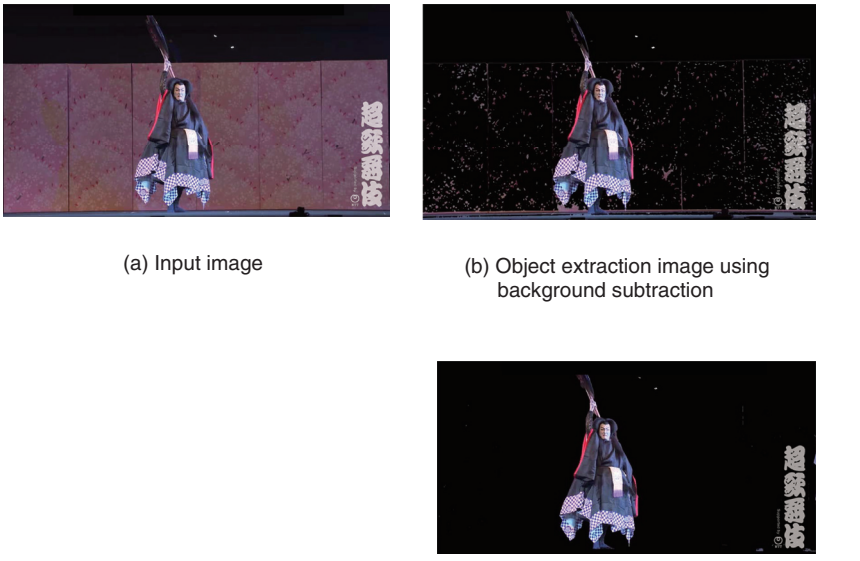


Fig. 5. Object extraction photography environment using RGB and IR cameras.



(a) Input image

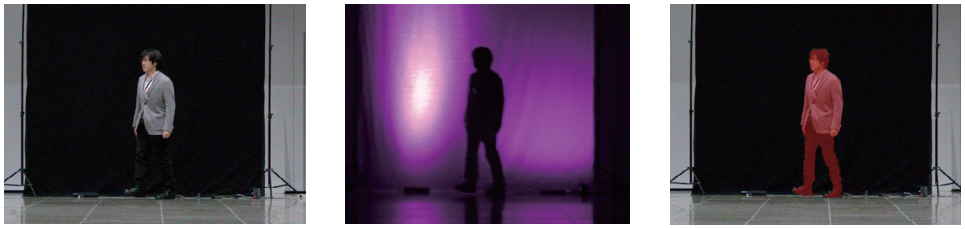
(b) Object extraction image using background subtraction

(c) Object extraction image using machine learning

Fig. 6. Real-time object extraction using machine learning on *kabuki* stage.

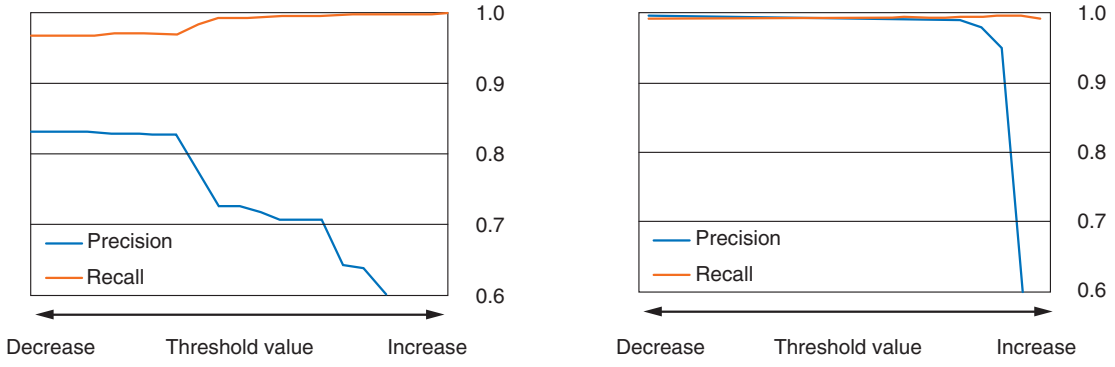
Hatsune Ono, played by virtual personality, Miku Hatsune, comes to its peak. Samples of the video at that point are shown in **Fig. 6**. The board behind Koretaka-shinno is being held by stage-hands, so it is not steady, and the image cannot be extracted well using background subtraction (Fig. 6(b)). We showed that it can be extracted more accurately using our system using machine learning (Fig. 6(c)).

We also checked the effect of using the IR camera (**Fig. 7**). We were able to confirm that objects can be extracted accurately, even in cases where it was difficult using an RGB camera. In this case, we did not do the automatic conversion of input feature vectors to a high order feature space or the threshold processing in the NN training, but we evaluated how robust the method using the IR camera would be by applying background subtraction and changing the threshold values (**Fig. 8**). We confirmed that using the IR camera



(a) Image using RGB camera (b) Image using IR camera (c) Result of object extraction

Fig. 7. Object extraction with machine learning using RGB and IR cameras.



(a) When using RGB camera only (b) When using both RGB and IR cameras

Fig. 8. Precision and recall using background subtraction.

enabled us to achieve very reproducible results as well as stable operation, even when the threshold values changed considerably.

We have introduced cases using an IR camera, but other characteristics suitable for extracting objects, depending on the conditions, can be input to the system, for example, depth maps generated using stereo cameras or LiDAR (light detection and ranging).

5. Future prospects

This article introduced a highly accurate object extraction method based on high-order feature values of objects, using machine learning to convert input feature vectors to a new feature space and performing the discrimination within an NN. We also introduced the use of features based on IR light, which is not visible to the naked eye, to handle use cases where this extraction is difficult using only an RGB camera.

In the future, to consider semantics when performing object extraction, we will work to perform object extraction using deep learning in real time and also study methods for extracting objects accurately when there are occlusions.

Reference

[1] H. Nagata, H. Miyashita, H. Kakinuma, and M. Yamaguchi, "Real-time Extraction of Objects with Arbitrary Backgrounds," NTT Technical Review, Vol. 15, No. 12, 2017.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201712fa7.html

Hirokazu Kakinuma
Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in information science from the Graduate School of Advanced Integration Science, Chiba University, in 2010. He joined NTT in 2010 and studied database management systems for Internet protocol television (IPTV). From 2013 to 2016, he worked at NTT Plala, where he developed set-top box web applications and released the Hikari TV 4K service for 4K digital TV. He moved to NTT in 2016 and is currently researching real-time image segmentation technology based on machine learning.

Jiro Nagao
Research Engineer, NTT Service Evolution Laboratories.

He received a Ph.D. in information science from Nagoya University, Aichi, in 2007. He joined NTT in 2007. His research interests include image processing for computer recognition and presentation.

Hiromu Miyashita
Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received a B.E. and M.E. from Keio University, Kanagawa, in 2008 and 2010. He joined NTT Service Evolution Laboratories in 2010, where he worked on human computer interaction, image and video processing, and ultrahigh-presence telecommunication services.

Yoshihide Tonomura
Senior Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.S. in electronics engineering from Nagaoka University of Technology, Niigata, in 2004 and a Ph.D. from Tokyo Metropolitan University in 2010. Since joining NTT, he has been researching and developing multimedia data transmission technologies. From 2011 to 2012, he was a visiting scientist at MIT Media Lab, Massachusetts, USA. He received the SUEMATSU-Yasuharu Award from the Institute of Electronics, Information and Communication Engineers in 2016 and the 26th TAF TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2011.

Hidenobu Nagata
Senior Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.

He received an M.E. in systems and information engineering from Hokkaido University in 2001. He joined NTT in 2001 and studied video technologies including video indexing and automatic summarization. From 2008 to 2014, he worked at NTT Electronics and developed transcoders and embedded audio IP. He is currently researching ultra-realistic communication technology including the immersive telepresence technology called “Kirari!”.

Kota Hidaka
Senior Research Engineer, Supervisor, Group Leader, NTT Service Evolution Laboratories.

He received an M.E. in applied physics from Kyushu University, Fukuoka, in 1998, and a Ph.D. in media and governance from Keio University, Tokyo, in 2009. He joined NTT in 1998. His research interests include speech signal processing, image processing, and immersive telepresence. He was a Senior Researcher at the Council for Science, Technology and Innovation, Cabinet Office, Government of Japan, from 2015 to 2017.