Real-time Extraction of Objects with Arbitrary Backgrounds

Hidenobu Nagata, Hiromu Miyashita, Hirokazu Kakinuma, and Mariko Yamaguchi

Abstract

With the development of Kirari! immersive telepresence technology, we strive to achieve a high sense of realism, presenting objects that are at a remote location as though they were right before the viewer. One Kirari! function making it possible to achieve this type of expression is arbitrary background real-time object extraction. This involves extracting objects from backgrounds such as a playing field or theater stage without requiring studio equipment such as a green screen. This article gives an overview of this technology and introduces further efforts toward increasing its speed, robustness, and detail.

Keywords: object extraction, high realism, immersive telepresence

1. Introduction

The objective of Kirari! immersive telepresence technology is to transmit an event happening at a venue to remote locations in near real time, presenting athletes or performers to viewers with extremely high realism so that they appear to be performing right in front of viewers. *Objects* such as athletes or performers are extracted from the video and transmitted in real time to the remote location, together with other information surrounding the objects and comprising the venue space such as the audience and other background video, sounds, and light levels. Then the space at the remote location is reconstructed from the data.

The objects are represented in life size, as though the performers themselves were actually there. To achieve this, the video itself must be very clear in order to give the impression that the athlete or actor is actually standing in front of the viewer, and the region containing the objects must be extracted and displayed very accurately. At NTT, we believe we can provide experiences that are more exciting and have greater impact if the audience feels the performance at the remote location is happening *right now* in front of them, rather than being prerecorded and processed video. For this reason, processing needs to be done in real time.

Usually in situations requiring real-time object extraction, a special photographic technique called chroma-keying is used. For chroma-keying, a screen in a single color such as blue or green is placed behind the object so that the difference in color between the object and the background screen can be used to extract the object. Our goal with the Kirari! immersive telepresence technology was to establish technology able to extract the object accurately and in real time from video taken of the space surrounding the object, whether it be a sports field or stage, rather than using a technique such as chroma-keying. This is the arbitrary background real-time object extraction that we are working on.

2. Overview of object extraction system

Work on the arbitrary background real-time object extraction technology is advancing in three directions: to increase speed, robustness, and detail. We discuss these directions below, but first we give an overview of the overall process used by the object extraction system and explain the context for the work in each of these directions.

An overview of the object extraction process is shown in **Fig. 1** [1]. The current object extraction

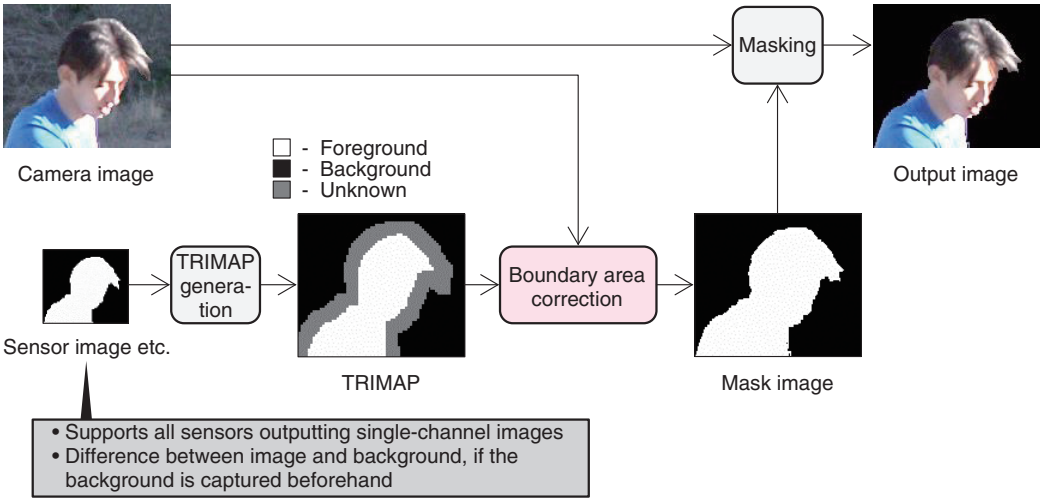

Fig. 1. Object extraction process.

system is based on the assumed use of sensors in addition to the imaging camera to capture the object. This is done to eliminate constraints on object extraction so that the range of applications can be expanded as much as possible in the future. As research on object extraction technology continues, further increases in the robustness of this process are being achieved.

Processing can be divided broadly into two stages. In the first stage, a sensing device and the background image (only if it can be captured ahead of time) are used to roughly identify the object area. In this stage, data (a trimap) are generated, and pixels are labeled in one of three categories: foreground, background, or unknown. Then, pixels in the trimap in the region labeled unknown (for which the foreground and background could not be distinguished during the rough identification stage and is assumed to contain the object boundary) are rapidly further classified as foreground and background, and corrections are applied to the boundary. Research continues to be done on this process, which is leading to increases in speed.

Finally, some objects targeted for extraction have an extremely complex boundary, such as with some traditional theater costumes. For these objects, the boundaries must be represented in fine detail to avoid degrading the sense of realism. In such cases, pixels in the unknown region are classified as foreground or background, but they are also assigned a transparency value called an α (alpha) value, which provides a more natural-looking boundary. The research and development (R&D) in this area is contributing to improvements that make it possible to achieve increased detail.

3. Three initiatives for technical development

Efforts are underway to improve the extraction technology. We describe here three improvements we hope to achieve and the initiatives being carried out to achieve them.

3.1 Initiatives for increasing speed

The boundary correction process used in the arbitrary boundary real-time object extraction technology is based on the assumption that the object will be extracted in detail, even when using low-resolution sensor images, and that the boundaries of the object area will be corrected using color data. The boundary correction process is based on a nearest-neighbor method, in which for each pixel labeled unknown in the trimap, the nearest (most similar) neighbor is found based on color data and coordinates, and the label of that pixel is copied to the pixel in question.

However, correcting this boundary is computationally intensive and becomes a bottleneck in producing the output video. Meanwhile, devices for capturing and displaying 4K/60 fps video are becoming more common, and there is increasing demand for Kirari! to support 4K/60 fps to provide an even higher sense of realism. However, processing high-resolution video captured using such equipment requires even more computation to perform boundary correction.

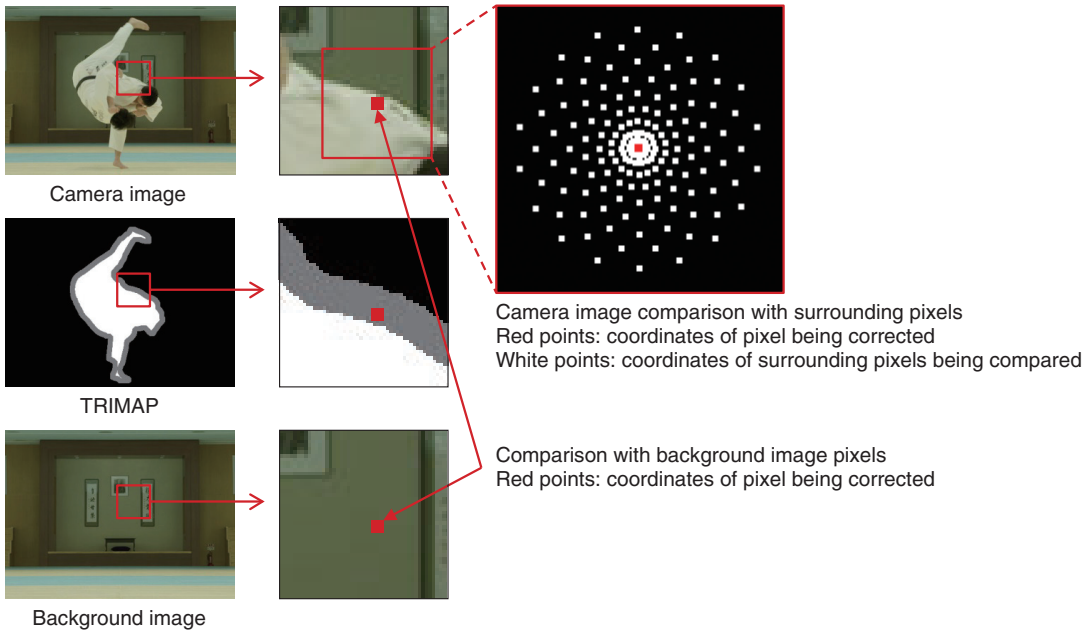

Fig. 2. Boundary correction.

As such, we set an objective to achieve real-time object extraction from 4K/60 fps video and are working to increase the speed of the boundary correction process, focusing in particular on reducing the amount of computation while also increasing the accuracy of extraction.

The boundary correction process has already been parallelized, but more technically, we can reduce the amount of computation by eliminating as many unnecessary comparisons as possible. To do so, we considered ways of searching for pixels that are *close enough* in terms of color data and distance between coordinates, rather than *the closest* in order to reduce computation while still obtaining a correct boundary.

When corrections are applied to the boundary, the pixel being corrected is compared with neighboring pixels. We adjusted this process so that comparisons with neighboring pixels are done more densely near the pixel and more sparsely further away. Comparisons are also done in order of increasing distance, meaning that the pixels that are nearest are compared first (**Fig. 2**). We also established thresholds and added a mechanism that stops doing comparisons and applies the label immediately if the distance derived from the color and coordinate data is below this threshold. In this way, *close enough* pixels are found quickly, and any further comparison operations are skipped.

However, the high resolution of 4K/60 fps video makes any error in extraction of the object noticeable, so higher accuracy than before is required. A particular weakness when generating the trimap with earlier techniques was that any error in labeling pixels as foreground or background was propagated to neighboring unknown pixels.

Therefore, we proposed the following method for obtaining the background image. For each unknown pixel, before performing any comparisons, we take the pixels at the correction coordinates in the input and background images. If the color information is close enough, we set the label of the pixel being corrected to background and avoid any comparison operations. This avoids propagation of any errors in background labels and makes it possible to obtain the correct object region [2].

3.2 Initiatives for increasing robustness

Objects to be extracted are not always in static environments or situations such as a studio that can be easily controlled. For example, the lighting and shadows behind the object can change during a sporting competition, and naturally, there are scenarios with spectators and other objects moving in the background. We are increasing the robustness of our technology to realize real-time object extraction that can also handle these sorts of situations.

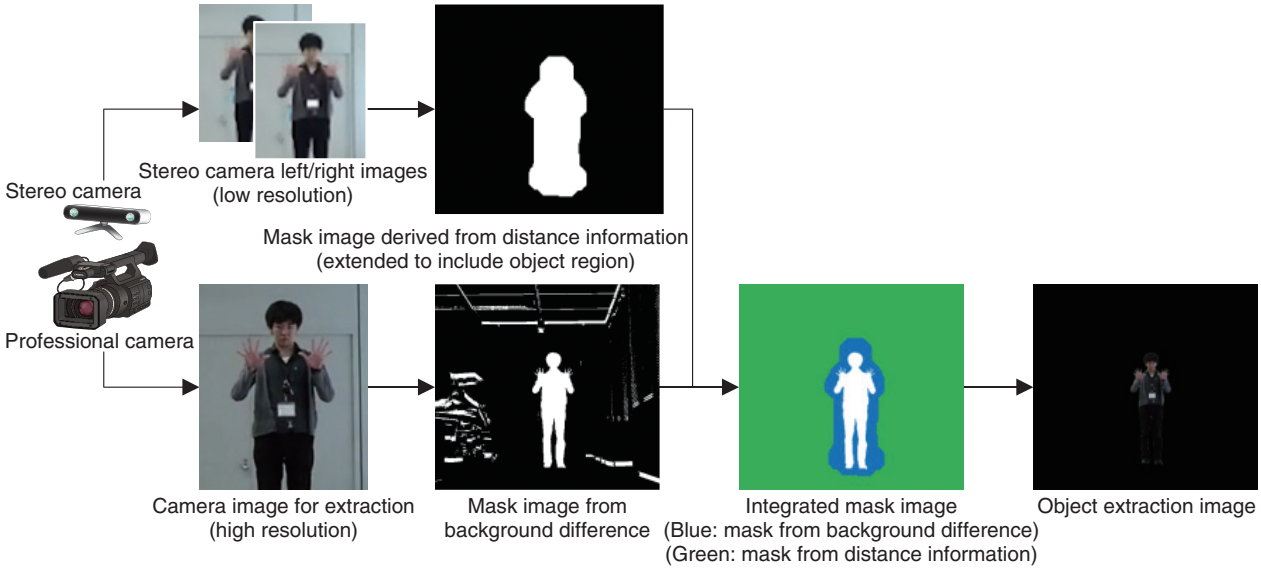

Fig. 3. Object extraction system incorporating a stereo camera.

At the NTT R&D Forum 2017 held in February, we built a system able to extract a specified object from full high-definition video in real time (30 fps or better) by combining video from a professional camera with partial depth information obtained using an off-the-shelf stereo camera (**Fig. 3**).

With this system, the system operator first selects an object to be extracted and specifies a rough range of distances for the object based on object distance information obtained using the stereo camera. Then, regions outside the specified depth range and regions with no change relative to the background are masked. This improves robustness for cases where the background changes, over just comparison with the background. This enabled the Kirari! system to extract one researcher from among several people in a meeting room, transmit the image, and project it onto the stage at the NTT R&D Forum 2017 in a remote location, where the researcher was able to interact with the moderator there.

However, although this system was able to roughly mask out undesired objects based on distance, it did not have sufficient accuracy to accurately separate the object border from the background (i.e., accuracy to correctly separate the object boundary when there are moving objects in the background from the camera's perspective). The system used an off-the-shelf stereo camera, so the mask image generated using distance data from the stereo camera was of a lower frame rate and had a different angle of view than the main professional camera image, from which the object was extracted. This resulted in frame mismatch between the stereo and main cameras when the object being extracted moved quickly or over a wide range, which decreased accuracy in extracting the object.

In the past, we have studied combinations of a camera used with various types of sensors such as time-of-flight and thermal sensors, but we found that images taken by all of these off-the-shelf sensors—just as with the stereo camera—differed from those taken by the main camera in resolution, field of view, or frame timing, and it was extremely difficult to calibrate and synchronize them perfectly, either spatially or temporally. The accuracy of measurements taken by sensor devices was also greatly affected by factors such as type of clothing, lighting, and measurable range. Making the object extraction system more versatile is therefore an issue to be addressed, and we intend to further improve the system to make it more robust in the future.

3.3 Initiatives for increasing detail

Kirari! is also desired for use in entertainment, for events such as concerts and *kabuki* theater. As such, we can envision scenarios when we would like to extract performers wearing complex and finely detailed costumes in traditional theater or a live concert. In such situations, the boundaries of extremely detailed hair styles or the textures of translucent costumes must be reproduced faithfully. This requires a

Fig. 4. Application of minimal matting process.

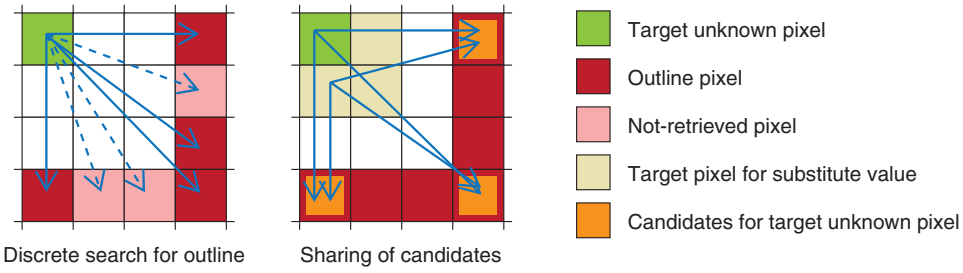

Fig. 5. Approximation in the matting process.

method for reproducing these complex and detailed boundaries that is more effective than the method of classifying pixels and applying corrections to the boundary.

We have studied a method that assigns a transparency value to pixels marked as unknown in the trimap in order to present extracted results naturally and in detail, even for objects with complex boundaries such as these. The method is called matting, and it is well known for naturally blending an extracted object into another image (background). Various algorithms have been proposed for matting, but algorithms to extract objects without the use of a green screen as in chroma-keying require large amounts of computing time and are difficult to use for real-time object extraction.

Because not all objects have complex boundaries, we only apply matting when necessary by classifying images according to edge intensity when the contrast with the background is low or the boundary region is very complex, as with some hairstyles [3]. For boundary regions where the object and boundary can be clearly distinguished, we apply boundary correction as in our earlier method (**Fig. 4**). This enables us to minimize the area that requires heavy computation, reducing the overall computation time.

We also introduced some approximation into the matting process. For matting, the transparency value applied to the pixel is derived from the surrounding pixels (candidate points), but candidate points are computed for each pixel handled, which increases the computation time. We therefore divided the matting region into small subregions and computed candidate points for each one, reducing the number of computations that need to be done. Note that within the small subregions, neighboring pixels have similar brightness and color, so computing transparency with this type of approximation enables extraction to be done without greatly degrading the quality of the boundary region (**Fig. 5**). We also introduced parallel processing using a graphics processing unit (GPU), which increased the processing speed by a factor of about 30 compared to our earlier implementation. The resulting process was used in the highlights section of the Kabuki Virtual Theatre demonstration held in Kumamoto, Japan, in March 2017. We will continue to study techniques for increasing speed in the future.

4. Future prospects

We have introduced efforts to improve our arbitrary background real-time object extraction technology, a core technology of Kirari!, in the three areas of speed, robustness, and detail. Regarding speed, we have achieved a 4K/60 fps processing speed while increasing accuracy. We have improved robustness with respect to background content and capture environment using an off-the-shelf stereo camera to measure

depth information. To increase detail, we produced an algorithm for processing complex boundaries that also minimizes any loss of speed. To this point, object extraction R&D has been advancing separately in the three approaches we have introduced in this article, but going forward, we plan to promote integration of this work, expansion of its applications through practical demonstration opportunities, and efforts toward practical application.

References

[1] H. Miyashita, K. Takeuchi, M. Yamaguchi, H. Nagata, and A. Ono, "Fast and Accurate Image Segmentation Technique Using Multiple Sensors," IEICE Tech. Rep., Vol. 116, No. 73, pp. 17–22, 2016.

[2] H. Miyashita, K. Takeuchi, H. Nagata, and A. Ono, "Fast Image Segmentation for 4K Real-time Video Streaming," IEICE Tech. Rep., Vol. 117, No. 73, pp. 189–190, 2017.

[3] M. Yamaguchi, H. Nagata, and A. Ono, "Adaptive Matting for Fast, Accurate Object Extraction," Proc. of 2017 IEICE General Conference, D-11-29, Nagoya, Aichi, Japan, Mar. 2017.

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 Internet protocol (IP). He is currently researching ultra-realistic communication technology including the immersive telepresence technology called "Kirari!".

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.

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 IP television (TV). 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.

Mariko Yamaguchi

Natural Communication Project, NTT Service Evolution Laboratories.

She received an M.E. from the Department of Computational Intelligence and System Science, Tokyo Institute of Technology Interdisciplinary Graduate School of Science and Engineering in 2015. She joined NTT in 2015 and is currently working on image processing technology for a virtual reality system.