Improving Depth-map Accuracy by Integrating Depth Estimation with Image Segmentation

Masato Ono, Yumi Kikuchi, Takashi Sano, and Shinji Fukatsu

Abstract

NTT Service Evolution Laboratories is conducting research and development on new media-processing technologies to enable natural interaction by presenting information that influences the visual senses. This article introduces our current research and development of a media-processing technology for improving the accuracy of depth maps by combining depth estimation with image segmentation and a system that uses this technology to generate HiddenStereo (a technology developed by NTT Communication Science Laboratories) images, which enables natural three-dimensional viewing from monocular two-dimensional images.

Keywords: 3D images, video streaming, depth maps

1. Integrating depth estimation with image segmentation to improve depth-map accuracy

Depth maps are a representation of the distance from the camera to the subject for each pixel in an image. They have various applications, including the familiar smartphone camera function that blurs distant background images when taking a picture and detecting nearby objects for self-driving vehicles.

We are conducting research and development (R&D) on a media-processing technology for improving the accuracy of depth maps by converting two-dimensional (2D) content into 3D content in addition to effectively representing newly created 3D content in the entertainment field (**Fig. 1**). This technology is composed of depth-map-generation technology, which generates accurate depth information from 2D images, and depth-map-optimization technology, which corrects depth maps for effective 3D representation and other purposes.

2. Technical points

(1) Depth-map-generation technology

There are various methods for obtaining depth maps, such as using the parallax between images from stereo cameras, combining camera images with information from a separate device (LiDAR*, etc.), or using various image-processing techniques. Deep learning has also been used recently to generate depth maps from 2D content, but the depth maps are generally of low resolution, and depth maps with clear outlines of the subjects cannot be obtained. When applied to recent high-definition (4K/8K) images, the resolution and quality of a depth map must also be high.

Regarding edge-preserving smoothing [1, 2], we are developing a technology for correcting depth maps, improving their accuracy, and enabling effective 3D representation by defining clear outlines for each subject in an image and performing edge-preserving smoothing on the results of image segmentation

* LiDAR: light detection and ranging.

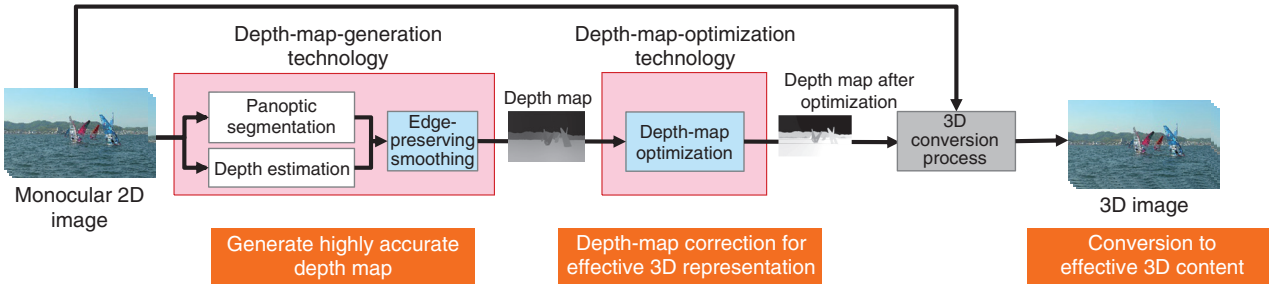

Fig. 1. R&D of media-processing technology for improving the accuracy of depth maps.

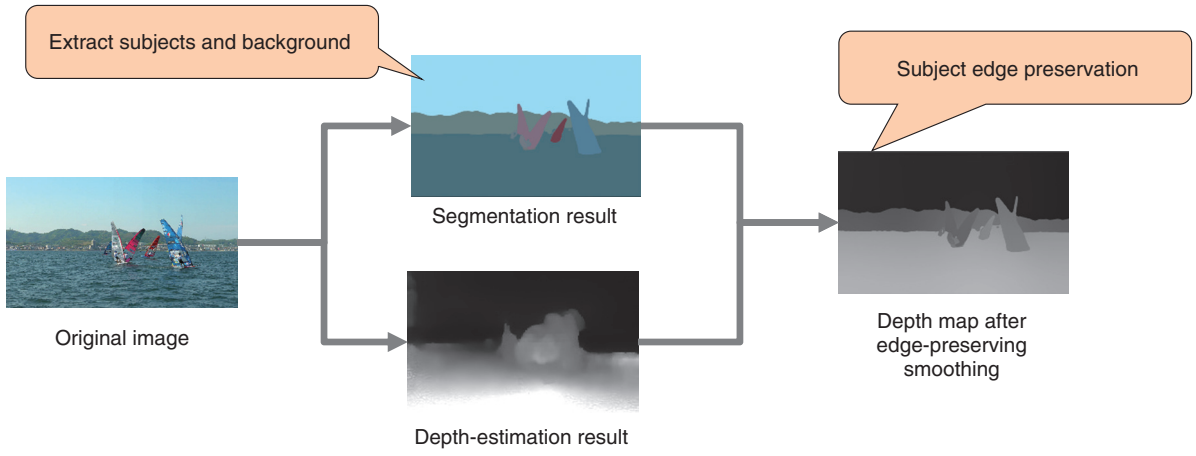

Fig. 2. Depth-map-generation technology.

(**Fig. 2**). Specifically, the resolution of a depth map can be increased by filtering the low-resolution depth map obtained though depth estimation using the segmentation-result image as a guide, while also preserving the edges of the subjects in the image. We use the results of *panoptic segmentation*, which is a combination of semantic segmentation and instance segmentation, to enable separation of subjects and the background.

With this technology, depth estimation and segmentation are only loosely related, so the algorithms for each can be adjusted as necessary. For example, we are currently using deep-learning algorithms for both depth estimation and segmentation, but other algorithms, such as those including self-supervised learning, can easily be substituted.

(2) Depth-map-optimization technology

The depth-map-generation technology described above can provide accurate depth maps, but when used as is, the maps may not necessarily be suitable for a particular application. For example, if the depth in the image is simply reduced to the depth range that can be expressed as a depth map, the depth difference in the image will not be sharp and the 3D expression will not be effective.

As a consequence, methods are used to correct depth maps to improve the sense of presence when viewing the 3D video. We are also developing a technology for optimizing depth maps, emphasizing the sense of depth surrounding the subject(s) that are the focus, to achieve a more effective 3D effect (**Fig. 3**). In particular, corrections are made from the following two perspectives.

(a) Limiting the range of depths used in 3D representations

A perceptual tendency when viewing in 3D is that it is difficult to distinguish fine depth differences in areas that are distant in the depth direction from the subject of focus (i.e., far in the background or close in the front), so we reduce the range of depth differences

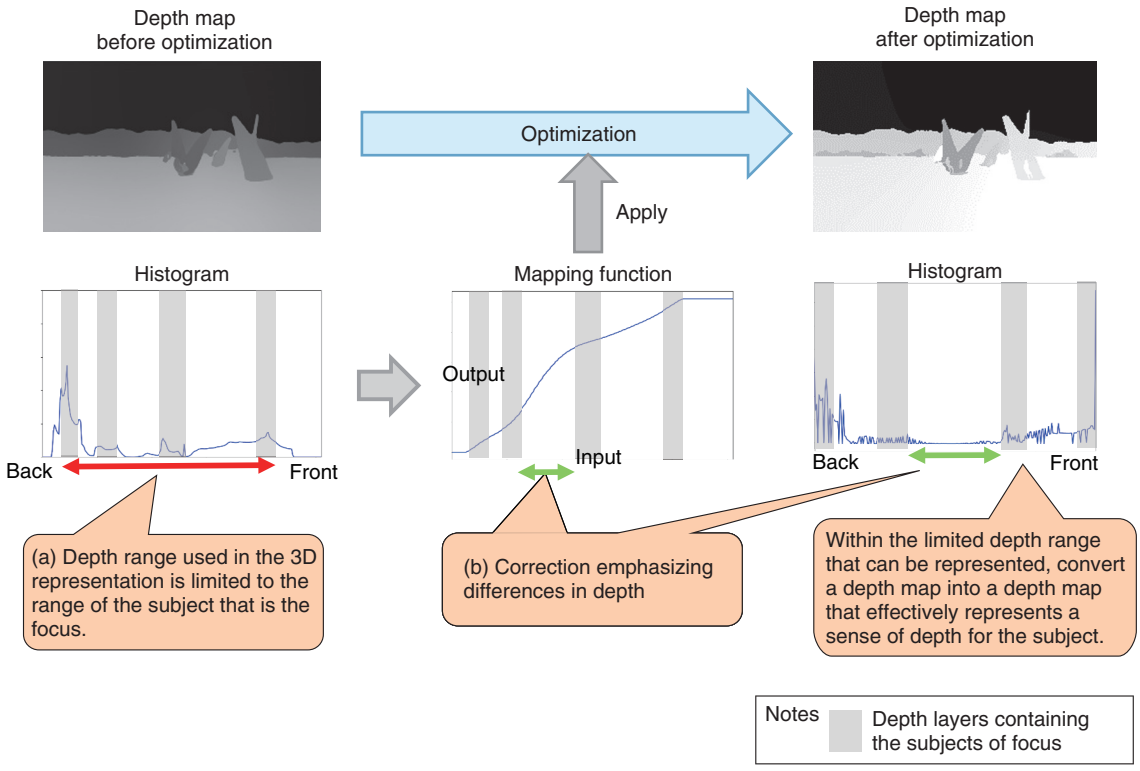

Fig. 3. Depth-map-optimization technology.

allocated to those depth ranges. Specifically, we analyze a depth map represented in steps from 0 to 255, create a histogram, and define ranges containing the subject as *effective depth ranges*. For example, this could be the range from the 5th percentile to the 95th percentile in the distribution of depth values from the depth map. Depths within this range are then extended over the range from 0 to 255, and values outside this range are mapped to either 0 (farthest) or 255 (nearest). Therefore, the 3D representation only expresses the effective depth range for 3D viewing.

(b) Representations emphasizing depth differences

By emphasizing the sense of depth in the range near the subject of focus, 3D video representing more subtleties can be generated. Specifically, from the depth-map histogram, we derive depth layers from the depth ranges containing the subjects of focus. The range of the depth layer with the subject to be emphasized most is then expanded, and the positions and depths of each of the depth layers are adjusted to emphasize the depth representation of that layer. This generates a mapping function for depth values, with input depth values on the horizontal axis and output values on the vertical axis, which has a segment for each depth layer. The segment for the layer containing the most emphasized subject has the greatest slope.

On the basis on the above two points, we derive a mapping function (nonlinear depth-map transform) to correct the depth map and apply it to the depth-map image, generating an optimized depth-map image.

There are various types of 2D archive video containing a range of camera work and scene changes, so the content of depth maps tends to change greatly with time. To enable depth-map optimization that can produce effective 3D representations for all such cases, we intend to continue applying our depth-map-optimization technology to various types of content, evaluate the results, and improve the quality of depth-map optimization.

3. Use in HiddenStereo

We are currently promoting the development of a system to generate HiddenStereo (a technology developed by NTT Communication Science Laboratories [3]) images, which enables natural 3D viewing using depth maps generated and optimized using all

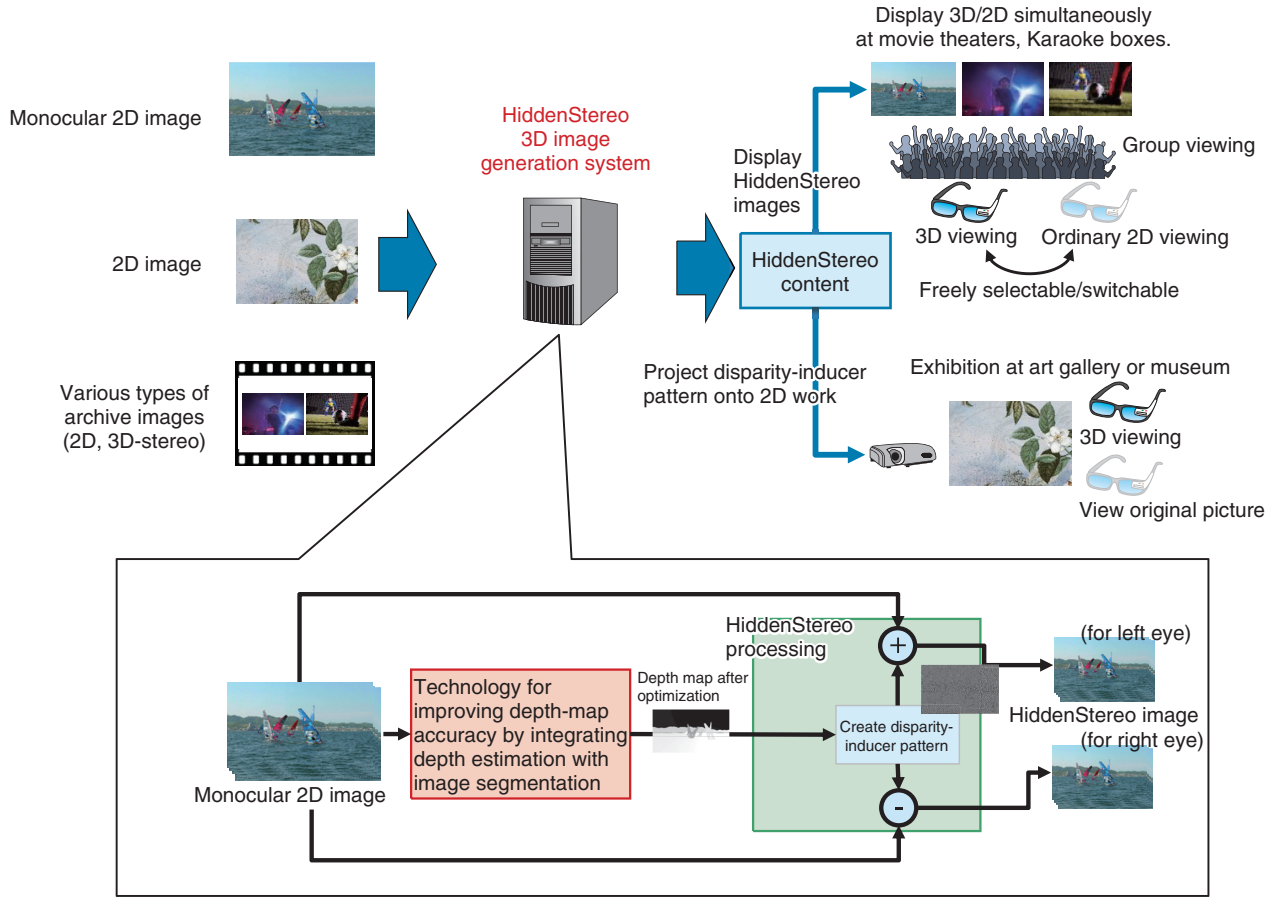

Fig. 4. HiddenStereo.

the above-mentioned technologies (**Fig. 4**).

HiddenStereo generates a disparity-inducer pattern from the original video and depth map, which provides depth information to human viewers. Images for the left and right eyes can be generated by adding or subtracting this pattern from the original image. When the left and right images are added together, the disparity-inducer patterns completely cancel each other out, leaving the original image, so that viewers without 3D glasses can see the 2D image clearly. Viewers with 3D glasses see the image with depth due to the effects of the disparity-inducer pattern.

This technology enables viewer-friendly 3D display, allowing each viewer to select how they want to view at any time, and all view the same display content. Since no special equipment is needed to display 2D and 3D at the same time, existing 3D display environments can be used as-is. Operational requirements for ordinary 3D video projection, such as presenting 3D and 2D video at different times, or preparing separate venues for 3D and 2D presentation are no longer necessary, which could reduce the operating costs for 3D video presentation.

By using our technology for improving depth-map accuracy, by integrating depth estimation with image segmentation, with HiddenStereo, we can convert not only new content captured with stereo cameras or created with 3D computer graphic production but also from 2D content produced in the past into 3D content.

4. Future developments

We introduced a media-processing technology we are developing for improving the accuracy of depth maps by integrating depth estimation with image segmentation, and a system using this technology to generate HiddenStereo images, enabling natural 3D viewing from monocular 2D image. We will continue investigation to further improve the speed and quality

of each of the component technologies through verification trials and other means to contribute to implementing businesses using natural interaction.

References

[1] K. He, J. Sun, and X. Tang, "Guided Image Filtering," 11th European Conference on Computer Vision (ECCV), Heraklion, Crete, Greece, Sept. 2010.

[2] G. Pestschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital Photography with Flash and No-Flash Image Pairs," ACM Trans. Graph., Vol. 23, No. 3, pp. 664–672, 2004.

[3] T. Fukiage, T. Kawabe, and S. Nishida, "Hiding of Phase-based Stereo Disparity for Ghost-free Viewing Without Glasses," ACM Trans. Graph., Vol. 36, No. 4, pp. 147:1–17, 2017 (Proc. of SIGGRAPH 2017, Los Angeles, USA).

Masato Ono
Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.
He received an M.E. in information engineering from University of Tsukuba, Ibaraki, in 2006. He joined NTT EAST the same year and engaged in network engineering, customer service management, and creating business-to-business services. He moved to NTT Service Evolution Laboratories in 2016 and is developing new services based on R&D products.

Yumi Kikuchi
Research Engineer, Natural Communication Project, NTT Service Evolution Laboratories.
She received a B.E. and M.E. in electrical engineering from Tokyo University of Science in 2000 and 2002. Since joining NTT in 2002, she has been researching and developing content-configuration methods and content-delivery systems. Her present research is focused on feature-point extraction of images and matching using the extracted feature points.

Takashi Sano
Senior Research Engineer, Innovative Service Research Project, NTT Service Evolution Laboratories.
He received a B.E. and M.E. in electronic engineering from Tokyo University of Science in 2002 and 2004. In 2004, he joined NTT Cyber Space Laboratories. He has been engaged in R&D on software and hardware design of video coding.

Shinji Fukatsu
Senior Research Engineer, Supervisor, Innovative Service Research Project, NTT Service Evolution Laboratories.
He received a Ph.D. in engineering from Osaka University in 2002 and joined NTT the same year. He has been engaged in R&D of human interfaces and video streaming services. He has also been engaged in planning and development of video streaming services at NTT Plala and in promoting standardization and international development of information and communication technology at the Ministry of Internal Affairs and Communications. He is currently engaged in R&D of remote-world infrastructure technology.