Development of a Semantic Segmentation System for Dynamic Occlusion Handling in Mixed Reality for Landscape Simulation

Daiki Kido[1]*, Tomohiro Fukuda*[2]*, Nobuyoshi Yabuki*[3]
[1,2,3]*Division and Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University*
[1]*kido@it.see.eng.osaka-u.ac.jp* [2,3]*{fukuda|yabuki}@see.eng.osaka-u.ac.jp*

The use of mixed reality (MR) for landscape simulation has attracted attention recently. MR can produce a realistic landscape simulation by merging a three-dimensional computer graphic (3DCG) model of a new building on a real space. One challenge with MR that remains to be tackled is occlusion. Properly handling occlusion is important for the understanding of the spatial relationship between physical and virtual objects. When the occlusion targets move or the target's shape changes, depth-based methods using a special camera have been applied for dynamic occlusion handling. However, these methods have a limitation of the distance to obtain depth information and are unsuitable for outdoor landscape simulation. This study focuses on a dynamic occlusion handling method for MR-based landscape simulation. We developed a real-time semantic segmentation system to perform dynamic occlusion handling. We designed this system for use in mobile devices with client-server communication for real-time semantic segmentation processing in mobile devices. Additionally, we used a normal monocular camera for practice use.

Keywords: *Mixed Reality, Dynamic occlusion handling, Semantic segmentation, Deep learning, Landscape simulation*

INTRODUCTION

Preserving good landscapes is important in enhancing our quality of life. To preserve good landscapes, it is necessary to predict and assess a planned landscape in the project. A landscape assessment means that a project executor hears various opinions from stakeholders and evaluates them (Fukuda, et al. 2014). Then it is difficult for stakeholders, especially non-experts like residents, to imagine the planned landscape in case three-dimensional (3D) objects have not existed yet. According to this context, the visualization of the planned landscape can help stakeholders to grasp the project and will preserve good landscapes.

The use of mixed reality (MR) for landscape simulation has attracted attention recently. MR merges real and virtual worlds (Milgram and Kishino, 1994), and can produce a realistic landscape simulation by overlaying a 3D computer graphic (CG) model of a building that does not yet exist on a camera view of a given area. Augmented reality (AR) is a similar concept that merges real and virtual worlds by overlaying computer-generated information on a real space; however, MR merges real and virtual worlds with high standards whereas AR only overlays information on a real space.

One challenge with MR that remains to be tackled is occlusion (i.e., the relationship between the physical and virtual worlds) (Shah et al., 2012). When physical objects occlude virtual objects, the augmented image may cause user confusion related to depth perception (Figure 1). Thus, it is crucial to handle occlusion appropriately in MR.

Previous studies of occlusion handling methods have been divided into three categories: model-based method, depth-based method, and contour-based method. Model-based methods create a 3DCG model of the real scene in pre-processing and compare the depth of the virtual objects with the 3DCG model to handle occlusion. Inoue et al. (2018) developed landscape design simulation system using MR. This system handled occlusion using a 3DCG model of the surrounding environment created with structure from motion (SfM) technology in pre-processing. In this system, the occlusion problem is solved by not rendering 3D model pixels that are obscured by the occlusion model. However, if a physical object such as vegetation changes its shape over time, it can be difficult to appropriately handle occlusion by changing the occlusion model's shape. Moreover, if the targets are moving objects such as car and person, it can be also difficult to handle occlusion. Depth based methods acquire the depth information of the real scene from a special camera in real-time and compare the depth of the virtual objects with the acquired depth information of the real scene to handle occlusion. Zhu et al. (2010) improved occlusion handling using depth information estimated by matching pixels from a stereo camera. Tian et al. (2015) converted the depth information acquired from RGB-D camera into a 3D model in real-time and handled occlusion. Some methods realized realistic occlusion handling by depth enhancement (Du et al., 2016, Holynski et al., 2018). However, depth-based methods have a limitation of the distance to obtain depth information and are unsuitable for outdoor landscape simulation. Contour-based methods detect and track the silhouette of the real objects and handle occlusion. Tian et al. (2010) obtained the contour of the specified occluding object by interactive object segmentation method, and tracked the object and handled occlusion. In this method, the occluding object needs to be displayed in the first frame, and segmented and tracked with high accuracy. Roxas et al. (2018) used semantic segmentation and a given depth map for occlusion handling. Semantic segmentation image pixels with a corresponding class of what is being represented. This method cannot handle dynamic occlusion because a given depth map is necessary, and also needs to use a high-end laptop computer in outdoor MR simulation because real-time semantic segmentation processing involves heavy processing.

This study focuses on a real-time dynamic occlusion handling method for MR-based landscape simulation. To this end, we developed a real-time

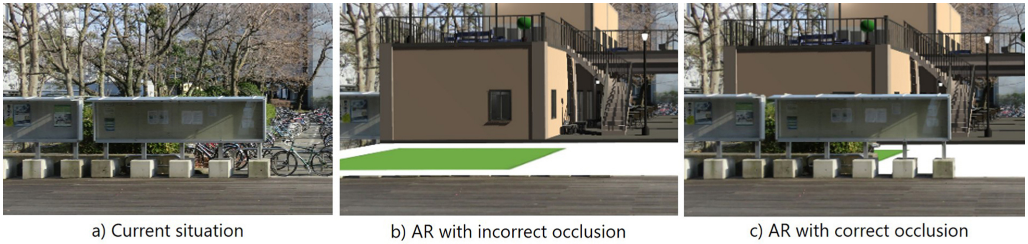

a) Current situation
b) AR with incorrect occlusion
c) AR with correct occlusion

Figure 1
Occlusion problem in MR (Inoue, et al. 2018)

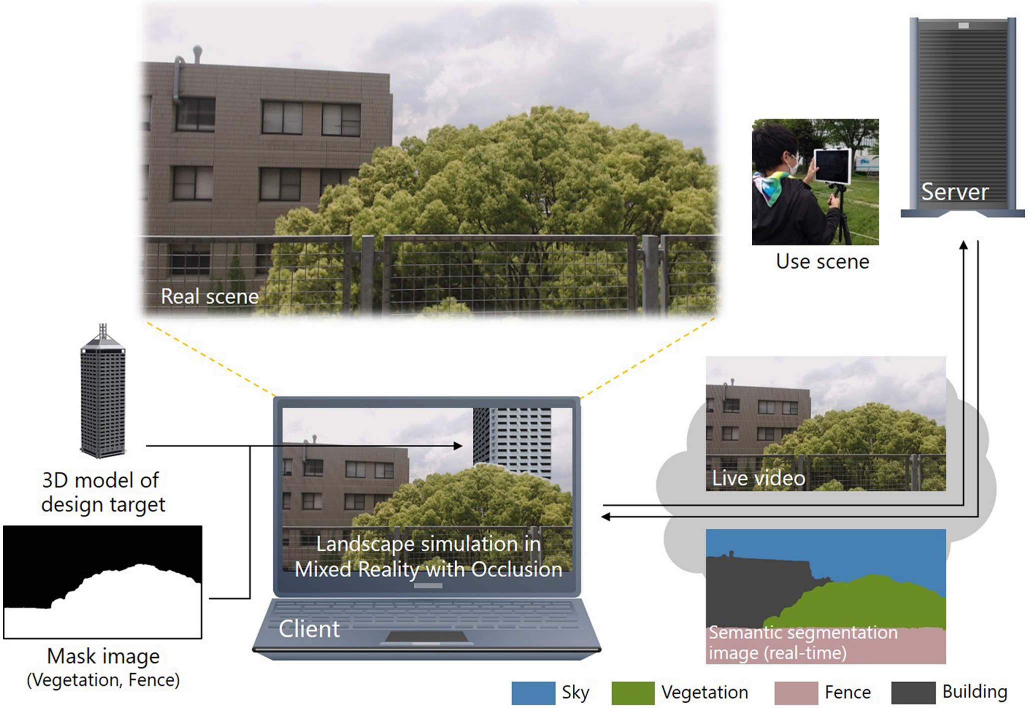
semantic segmentation system to perform dynamic occlusion handling. Our proposed method can extract each object's region and handle occlusion dynamically in outdoor MR simulation. Thus our proposed system can implement more realistic landscape simulation with the expression of the relationship between a 3DCG model and physical objects including moving objects such as car and person and help stakeholders to understand a planned landscape more correctly. We designed this system for use in mobile devices such as standard laptop computers and tablets, thus, it enables client-server communication for real-time semantic segmentation processing in mobile devices. Additionally, we used a monocular camera for this system.

DEVELOPMENT OF A SEMANTIC SEGMENTATION SYSTEM FOR DYNAMIC OCCLUSION HANDLING

Overview of our proposed system is shown in Figure 2. As explained in Chapter 1, we adopted a semantic segmentation technique to handle occlusion dynamically in outdoor MR simulation. Semantic segmentation can extract each object's region frame by frame, thus, it is considered that dynamic occlusion handling can be realized using this technique.

Real-time semantic segmentation processing involves heavy processing and thus requires a high-end desktop computer. However, our proposed system is to be used for landscape simulation outdoors, so it needs to be used on a mobile device. We there-

Figure 2
Overview of our proposed system


Real scene
Use scene
Server
3D model of design target
Live video
Landscape simulation in Mixed Reality with Occlusion
Client
Mask image
(Vegetation, Fence)
Semantic segmentation image (real-time)
Sky
Vegetation
Fence
Building

fore developed a system in which a client device transfers image frames to a server, which performs semantic segmentation processing on those frames and sends processed frames back to the client. Moreover, many semantic segmentation modules have been proposed due to the advancement of computer vision recently. Thus, we designed this system whose semantic segmentation module can be replaced when more sophisticated semantic segmentation module is developed. We used the Unity game engine to develop our proposed system because Unity game engine can implement a Graphic User Interface (GUI) based MR system.

Client-server communication

For client-server communication, we used a Python web application framework known as Flask, which supports the development of web-based services. When a client device connects to the web application on the server it sends an HTTP request, which the server receives along with an image from a camera on the client device. The server then performs semantic segmentation processing and transfers the segmentation image back to the client device as an HTTP response (Figure 3).

Next, we displayed the frames acquired from the client device camera and the segmentation images acquired from the server in the Unity engine on the client side. We used Unity's www class to receive segmentation images from the web application on the server. If the frames acquired from the camera are displayed as is, there would be a gap between those frames and the corresponding segmentation images because of the latency of the client-server communication. Therefore, we delayed and adjusted the display of client-acquired frames to the display of segmentation images, solving this mismatch because was assumed it would lead to inappropriate occlusion handling.

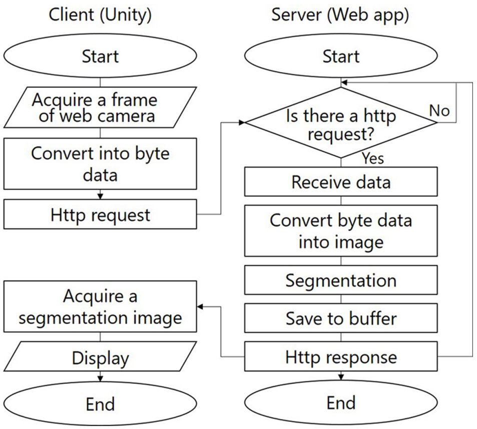


Figure 3
Flow chart of the client-server communication

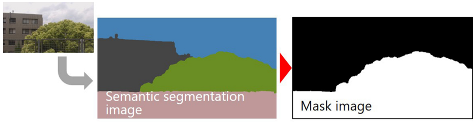


Figure 4
Mask image creation (targets: vegetation, fence)

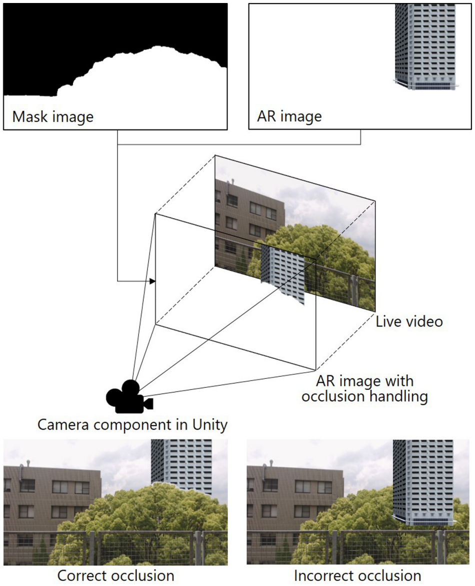


Figure 5
Occlusion handling by not rendering AR image pixels that merged into a mask image

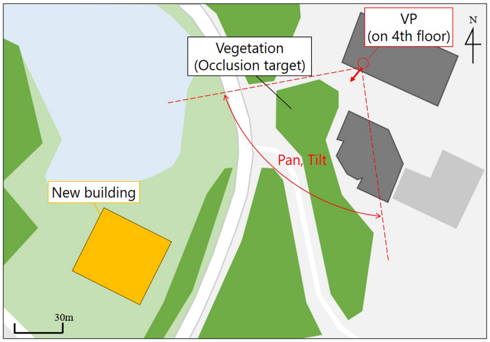


Figure 6
Arrangement of the newly constructed building

Occlusion handling

To implement dynamic occlusion handling, we must create a mask image of the occlusion handing target object. We created a mask image with RGB values in the segmentation image using OpenCV for Unity (ver. 2.2.8), which is a Unity image processing plugin (Figure 4). We then defined an occlusion handling target object and created the object's mask image. We then merged the mask image into an image of a 3DCG model, referred to as an AR image, and handled occlusion by not rendering AR image pixels that merged into the mask image (Figure 5).

VALIDATION EXPERIMENT

We validated our semantic segmentation system for dynamic occlusion handling. We adopted ICNet (Zhao et al., 2018), which is a semantic segmentation method based on deep learning, because of its processing speed and accuracy. We applied the Cityscapes dataset (Cordts et al., 2016), which was created for understanding urban scenes as ICNet's semantic segmentation dataset. We used a laptop computer with Intel Core i5-8250U of CPU, 8GB of RAM, and Intel UHD Graphics 620 of GPU as a client device and a desktop computer with Intel Core i7-8700K of CPU, 32GB of RAM, and NVIDIA GeForce GTX 1080Ti 11GB of GPU as a server device. A Logicool HD Pro Stream Webcam C920 was also used. We acquired frames with 1024 × 576 pixel resolution from a webcam connected to the laptop and conducted client-server communication over a local area network (LAN) as this was the only possible environment.

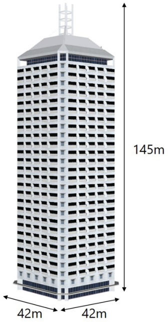


Figure 7
Measurement of the newly constructed building model

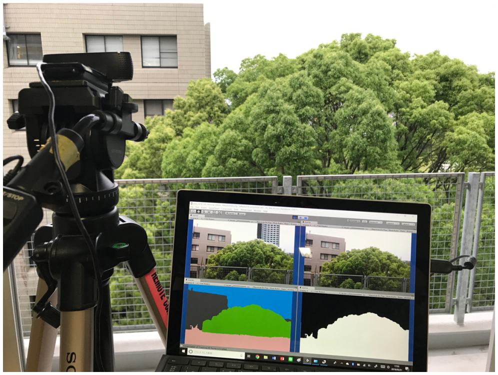


Figure 8
Experimental photograph

Outdoor MR simulation was conducted to validate the occlusion handling of the proposed system. Figure 6 shows the arrangement of the new building and the camera's position and orientation. Figure 7 shows the measurement of the new building model. Then, the viewpoint was on the fourth floor of the building because the server device was equipped on the fourth floor of the building and client-server communication was conducted over LAN (Figure 8). Veg-

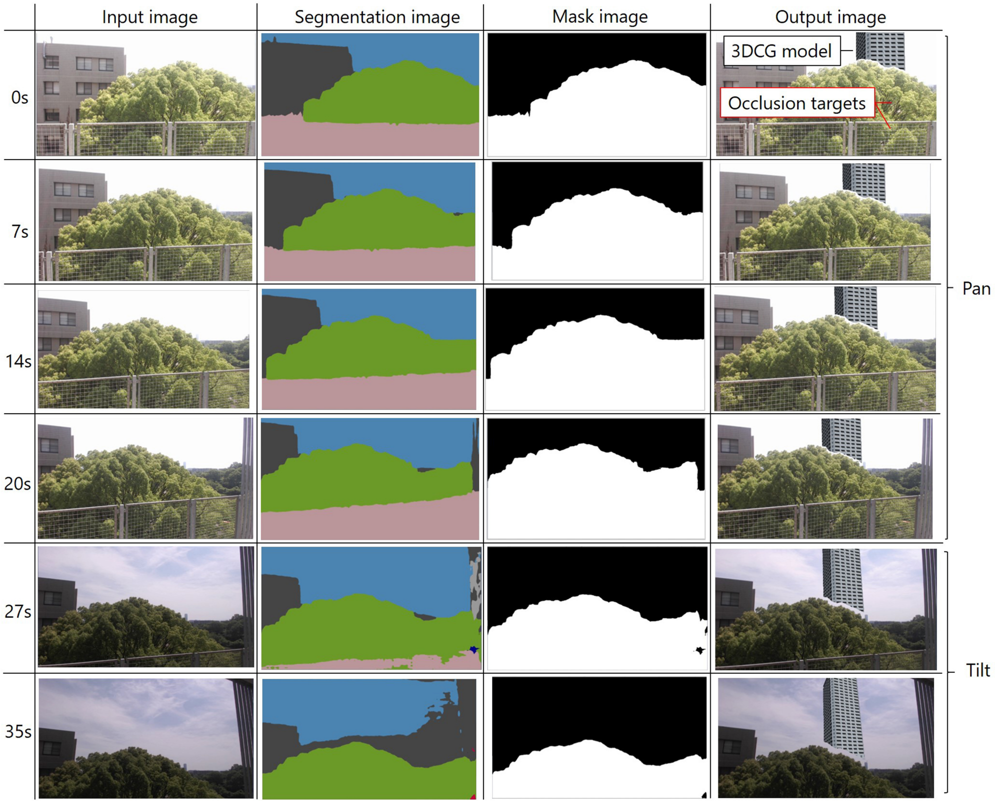


Figure 9
Validation results of dynamic occlusion handling

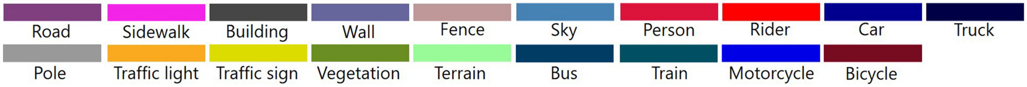


Figure 10
Labels which can be detected in this experiment

etation and fence equipped on the fourth floor of the building were defined as occlusion targets in this experiment. The webcam was panned in a horizontal direction in the range of the red dotted line of Figure 6 during 20 seconds. After 20 seconds, the webcam was tilted up during 15 seconds. Camera's position and orientation were computed using the method which solves the perspective n-points (PnP) problem and eliminates the outliers using the random sample consensus (RANSAC) method proposed by Inoue et al. (2018). The validation results are shown in Figure 9. Labels and objects which can be detected in this experiment are shown in Figure 10.

From this validation, we confirmed that our system can create semantic segmentation images and mask images using frames acquired from a webcam, thus handling appropriate dynamic occlusion in MR. This system also achieved dynamic occlusion handling using a general mobile device, meaning that this system can be used widely and easily for landscape simulation. However, the processing speed was about 5 frames per second (fps). Processing speed that displays only frames acquired from the webcam was about 30fps, and the processing speed of semantic segmentation with ICNet was about 65 fps. Therefore, it was confirmed that the communication speed between the client-server should be improved. And also it was confirmed that the result of semantic segmentation processing contained false extraction, especially in the result after 35 seconds. The accuracy of semantic segmentation processing should be validated.

CONCLUSION

This research accomplished the following:

- We developed a system which enables dynamic occlusion handling in MR using image processing techniques that create a mask image from a segmentation image and merge that mask image into an AR image.
- We developed a system in which a client device acquires frames and transfers them to a server that implements semantic segmentation processing on the frames and transfers them back to the client device.

Future work should adapt this system to a wide area network (WAN) and Internet environment. Additionally, we should validate our proposed system at an outdoor venue like an architectural project, and also validate in the case where occlusion targets are moving objects such as car and person.

ACKNOWLEDGEMENTS

This research has been partly supported by JSPS KAKENHI Grant Number JP19K12681.

REFERENCES

Cordts, M., Omran, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. 2016 'The cityscapes dataset for semantic urban scene understanding', *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pp. 3213-3223

Du, C., Chen, Y., Ye, M. and Ren, L. 2016 'Edge snapping-based depth enhancement for dynamic occlusion handling in augmented reality', *The 15th IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2016)*, pp. 54-62

Fukuda, T., Zhang, T. and Yabuki, N. 2014, 'Improvement of registration accuracy of a handheld augmented reality system for urban landscape simulation', *Frontiers of Architectural Research*, 3, pp. 386-397

Holynski, A. and Kopt, J. 2018, 'Fast depth densification for occlusion-aware augmented reality', *ACM Transactions on Graphics (TOG)*, 37 (6), pp. 1-11, 194

Inoue, K., Fukuda, T., Cao, R. and Yabuki, N. 2018 'Tracking Robustness and Green View Index Estimation of Augmented and Diminished Reality for Environmental Design: PhotoAR+DR2017 project', *Proceedings of the 23rd International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA 2018)*, pp. 339-348

Milgram, P. and Kishino, F. 1994, 'A Taxonomy of Mixed Reality Visual Displays', *IEIXE Transactions on Information and Systems*, E77-D, 2, pp. 1321-1329

Roxas, M., Hori, T., Fukiage, T., Okamoto, Y. and Oishi, T. 2018 'Occlusion Handling using Semantic Segmentation and Visibility-Based Rendering for Mixed Reality', *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology (VRST 2018)*, pp. 1-8, 20

Shah, MM., Arshad, H. and Sulaiman, R. 2012 'Occlusion in augmented reality', *Proceedings of the 8th International Conference on Information Science and Digital Content Technology (ICIDT 2012)*, pp. 372-378

Tian, Y., Guan, T. and Wang, C. 2010, 'Real-Time Occlusion Handling in Augmented Reality Based on an Object Tracking Approach', *Sensors*, 10 (4), pp. 2885-2900

Tian, Y., Long, Y., Xia, D., Yao, H. and Zhang, J. 2015, 'Handling occlusions in augmented reality based on 3D reconstruction method', *Neurocomputing*, 156, pp. 96-104

Zhao, H., Qi, X., Shen, A. and Jia, J. 2018 'ICNet for Real-Time Semantic Segmentation on High-Resolution Images', *Proceedings of European Conference on Computer Vision (ECCV 2018)*, pp. 418-434

Zhu, J., Pan, Z., Sun, C. and Chen, W. 2010, 'Handling occlusions in video-based augmented reality using depth information', *Computer Animation and Virtual Worlds*, 21, pp. 509-521