2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

3D Reconstruction Based on Grouping Similar Structures for Images Acquired in the Fukushima Daiichi Nuclear Power Station

Takashi Iamabuchi, *Member, IEEE*, Toshihide Hanari, Kuniaki Kawabata, *Senior Member, IEEE*

Abstract— **This paper describes a 3D reconstruction based on grouping similar structures for the aim of generating 3D information for understanding the workspace from images acquired inside the Primary Containment Vessel (PCV) of the Fukushima Daiichi Nuclear Power Station. In the decommissioning works, preliminary surveys are carried out in the PCV, and the workers need to understand the workspace from a large number of camera images, which requires a great deal of effort. We are currently working on 3D reconstruction from PCV camera images; however, one of the challenges is to improve the visibility of the reconstructed model containing noise and artifacts. In this study, we propose a method of grouping similar structures on images and utilizing predicted group labels for 3D reconstruction process to highlight structures shapes and to refine 3D modeling. Our key idea is to perform unsupervised segmentation for grouping similar structures that are suitable for images acquired in the PCV because they are difficult to assign correct semantics for unclear structures and the few learning resources. We show on the reasonable performance the proposed method by validating it using video images of a typical plant environment and survey videos of the PCV taken under adverse conditions, such as radiation noise.**

I. Introduction

In the decommissioning of the Fukushima Daiichi Nuclear Power Station (FDNPS), preliminary surveys have been carried out to investigate the conditions inside the Primary Containment Vessel (PCV) to consider the construction method of the remaining fuel debris removal from the vessel [1]. The survey of Unit 1 using an underwater Remotely Operated Vehicle (ROV) revealed what appeared to be fuel debris, exposed reinforcing steel, and skeletons of damaged structures on video images acquired by camera sensors. In these surveys, the conditions of the workspace are elucidated through visual inspection based on several hours of video images captured in a single day, which is a huge burden for workers. In addition, the difficulty to in ascertaining the geometric shape of structures in 3D space from a 2D video image makes it difficult to understand the condition. Hereafter, the use of remotely operated robots will accelerate investigations inside the vessel in preparation for full-scale fuel debris removal. A technology that reconstructs a 3D scene from only camera images as quickly and precisely as possible will improve the overall efficiency of decommissioning work.

Therefore, we have been developing a method to perform 3D reconstruction from video images obtained from surveys in the PCV [2]. Preliminary analysis of video scenes captured by ROV showed that several 3D scenes, as 3D point cloud data, could be reconstructed by applying Structure from Motion (SfM) and Multi-View-Stereo (MVS) [3-4]. Here, the camera trajectory estimation and 3D reconstruction failed in several cases due to missing key-points caused by indistinct structures on the images due to insufficient lighting conditions, turbid water, and dust or steam in the air. A further factor is the failure of feature matching between images caused by radiation-affected imaging noise and floating objects in water. To address these issues, we employed key-points extraction and feature matching methods applying deep learning to extract as much information as possible from a small number of valuable data [5]. These methods enabled a wide range of reconstruction by ensuring many feature matching results; however, the results had many noisy and undesirable reconstruction points resulting from feature mismatching, which are difficult to avoid by parameter adjustment. Such unreliable feature matching results lead to reduced visibility of the 3D reconstruction result. Furthermore, this prevents the separating parts and causes the generation of unnecessary surfaces when performing the 3D modelling required for understanding the conditions, radiation calculations, and waste volume estimation in decommissioning works. To improve visibility and refine the 3D model, it is necessary to extract the shape of structures from the 3D reconstruction results, eliminating unnecessary points. In other words, we need to separate the 3D reconstruction results by the structure shapes and make them selectable.

One of the possible solutions is to segment the region of the 3D point cloud. We previously confirmed that it is possible to separate and utilize 3D point cloud by using semantic segmentation deep learning models trained by typical plant structure categories [6]. However, in the current case, there are two major issues when applying supervised learning via assignment semantics to 3D point cloud. The first issue is that the training resources (i.e., the reconstructed 3D point cloud) are noisy and insufficient quality data, and in addition, we cannot ensure enough quantity. Another is that the target object is too uncertain to assign the ground truth semantic labels. The structures projected on images are still in the survey and analysis phases, and their origin is not known. Our idea in this work is to apply an unsupervised learning approach to 2D images. This is because, for our purposes, we do not need semantics; we only need to know the shape of structures. By introducing an unsupervised learning model, we expect to segment the shape of structures in an image and automatically estimate classes with similar features (hereafter, grouping).

All authors are with the Collaborative Laboratories for Advanced Decommissioning Science (CLADS), Sector for Fukushima Research & Development, Japan Atomic Energy Agency (JAEA), Fukushima, Japan (phone: (+81)240-26-1188; e-mail: imabuchi.takashi@jaea.go.jp).

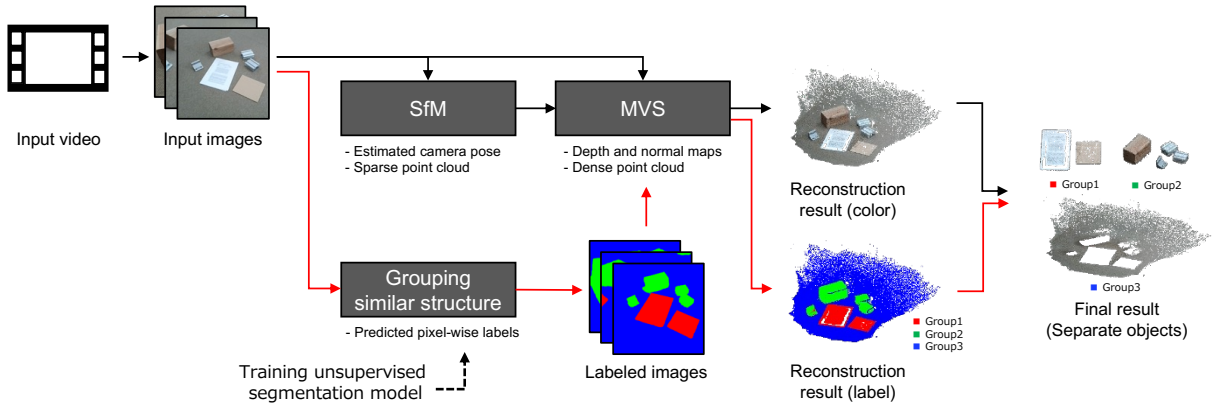


Figure 1. Overview of 3D reconstruction based on grouping similar structures.

In this study, we propose a 3D reconstruction method based on grouping similar structures: groups are trained by unsupervised segmentation of the image to highlight the structure shapes and eliminate the noise-causing regions in the survey video of the PCV. The grouping results are used in the 3D reconstruction process, which finally generates a 3D point cloud segmented by grouping labels. In addition, grouping results can help refine the 3D modeling process. In our verification, we show that the proposed method works reasonably using videos captured in a typical plant environment, and its experimental application to publicly available survey videos obtained at the PCV of the FDNPS.

II. 3D Reconstruction Based on Grouping Similar Structures

In this section, we describe 3D reconstruction method based on grouping similar structures. Fig.1 shows an overview of the proposed method. Our method is a pipeline that applies a grouping module using a pretrained unsupervised deep learning model to 3D reconstruction by SfM and MVS. The predicted group labels are incorporated in the MVS process when merging stereo image pairs to separate the 3D reconstruction results. In the following, we describe the details of our methodology.

A. Grouping Similar Structures

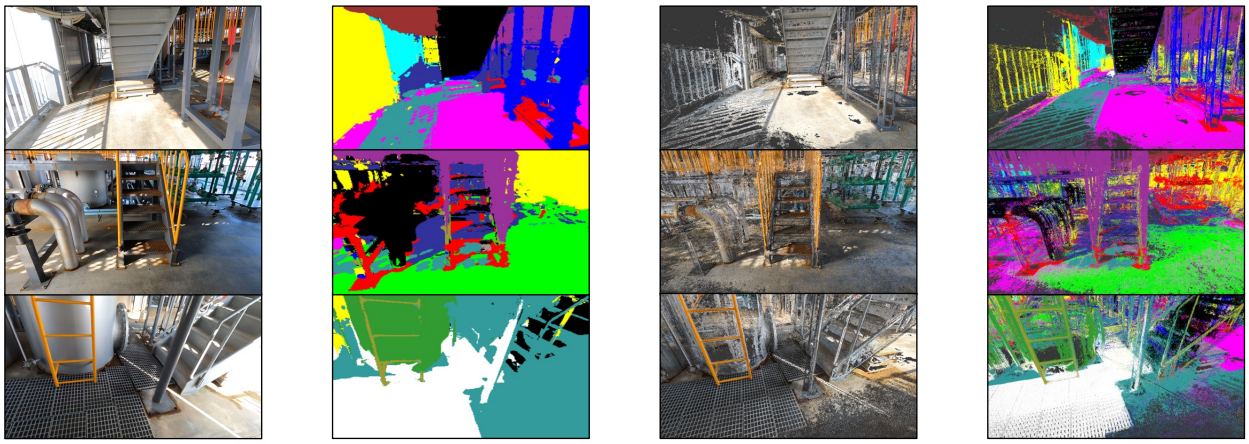
The segmentation task by deep learning generally requires a sufficient amount of training data; however, survey videos of the PCV are uncertain and not suitable for annotating semantics. Here, we attempt to group similar structures without the need for annotation by applying an unsupervised segmentation model. We employ STEGO as an unsupervised segmentation model [7]. STEGO calculates the final segment result using clusters created in the feature space by simultaneously learning the similarity of embedded features and segmentation features between self-, random-, and similar images. DINO as a model of feature distillation approaches and a multi-layer perceptron are used as the backbone and segmentation head, respectively [8]. Feature distillation refers to the transfer of knowledge acquired in one model to another model. DINO employs self-supervised learning and argued that using VisionTransformer as feature embedding, it is possible to segment the major objects in an image in the ImageNet dataset. Thus, we can expect to be able to segment the major structural shape and background (e.g., region of indistinct structures on the images by turbid water) in this study as well. Furthermore, we expect that as a side effect, the model will be able to predict noise and floating objects as a group. The other reasons for the model selection include end-to-end learning and highly accurate segments. Although Segment Anything Model [9] and other extension models have appeared in recent years, they target interactive segmentation and are not suitable for an automated approaches such as the one used in this study.

In the training phase, the STEGO model is trained using images output from the videos. By training multiple videos together, similar structures in various scenes can be trained as a group. In the prediction phase, group labels are predicted as pixel-wise labels based on a pretrained model by forwarding the time-series images output from the video. Normally, a trained model is used to predict unknown pattern data; however, in this study, we use the same training and prediction sources to predict group labels from input videos.

B. Reconstruction Based on Predicted Group Labels

The predicted group labels are used in the MVS process in the 3D reconstruction pipeline, which assigns label information to the 3D point cloud. In this study, we use the hloc framework for the 3D reconstruction pipeline [5]. The hloc is based on a 3D reconstruction method by SfM and MVS with deep learning models. First, image feature points are extracted from a set of input images, and then correspondence point matching is performed between pairs of images. Here, we used SuperPoint and SuperGlue [10-11], which are methods that apply deep learning. Based on both the information, the camera trajectory and the reconstructed points are estimated based on the triangulation principle. Up to this sparse stereo reconstruction, the process is the same as that of general pipelines. MVS is the postprocessing of SfM to reconstruct dense point cloud. The hloc employed the MVS process based on COLMAP using a patch matching method with sparse points to perform dense 3D reconstruction. During this process, the depth and normal maps are estimated from each image while optimizing the patches. Finally, the colored 3D point cloud are reconstructed from the estimated camera pose using the patch information and the pixel information from the input image. In addition to color information, we

Figure 2. Example of video frame of typical plant environment data.



(a) Input image (b) Grouping labels (c) 3D reconstruction result (d) 3D reconstruction result (grouping labels)

Figure 3. Result of grouping 3D reconstruction for typical plant environment data.

assign the predicted group label to each reconstructed point. In COLMAP, the function *stereo_fusion* corresponds to the color point reconstruction process, which we modified. This modification makes it possible to display and hide each label in the point cloud for applications. Furthermore, 3D point classification methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can be used to separate instances automatically within each group [12].

III. VERIFICATION

In this section, we verify that the proposed method works as expected on the two datasets. The first target is a clean video image taken in a typical plant environment to confirm the performance of grouping for various types of structural geometries. We also confirm the reconstruction result by grouping information and its processing time. The second target is a publicly available survey video of the PCV. We applied the proposed method experimentally to images taken under relatively noisy and insufficient lighting conditions. We will then confirm whether visibility can be improved by highlighting structures in groups. The results of the two videos are presented in the following sections.

A. Result for Plant Environment Video

We performed 3D reconstruction based on grouping similar structure on video images of a typical plant environment to evaluate the performance of our method. The target is a mockup plant of the Fukushima Robot Test Field, which is a robotic test environment built after the FDNPS accident to promote decommissioning research [13]. The video was captured by GoPro 7, and the cameraman went around the first floor of the mockup plant for two laps with changing angles. Fig.2 shows an example of a captured video frame. The video was approximately 3.4 minutes long, and the total number of output frames by 15 fps was 3,092 with a resolution of 1920×1080 pixel. For training the model, the number of batches was set to 16 and the number of training steps was set to 5,000. The target classes were separated into 15 classes, and the backbone for feature distillation was ViT-b/8. The total number of data was 15,460, because the data were created by pre-cropping images into 5 segments strategy. All frames were used to train the STEGO model. The time required to perform each processing step was measured programmatically. The specifications of the computer used for processing were CPU: Intel Xeon W-2235 3.80 GHz, GPU: Nvidia GeForce RTX3090 24 GB, memory: 64 GB.

TABLE I. RESULT OF PREDICTED GROUPS BY STEGO FOR THE VIDEO OF TYPICAL PLANT ENVIRONMET

No.	Color	Group contents	Acc.
1		Pump (H), pipe (H), support (H), bulb	74.85%
2		Floor	98.79%
3		Pipe (V), support (V)	83.12%
4		Handrail (metric), pump (V), outside	51.03%
5		Wall, door	90.16%
6		Floor	96.48%
7		Frame, support (H), light fitting	60.46%
8		Tank	53.65%
9		Floor, frame	60.86%
10		Pipe (V), support (V), ladder	43.48%
11		Floor, objects away from the camera	81.29%
12		Handrail, fire suppression, outside	75.92%
13		Pipe, tank, stair	88.30%
14		Checkered steel plate, rusty structure	93.37%
15		Grating	88.58%

(H): Horizonal, (V): Vertical

Fig.3 shows the results of 3D reconstruction using the proposed method. Fig.3 (a) shows the input image and Fig.3 (b) shows the result of the prediction by STEGO. Fig.3 (c) shows the 3D reconstruction and Fig.3 (d) shows the result of group labeling. The results show that for each input image, the structural shape region is well segmented. Table 1 shows the results of reviewing all images and confirming the structures contained in the 15 groups. The table shows that major structures such as pipes, supports, tanks, gratings, and frames can be predicted as separate groups. Then, as shown in

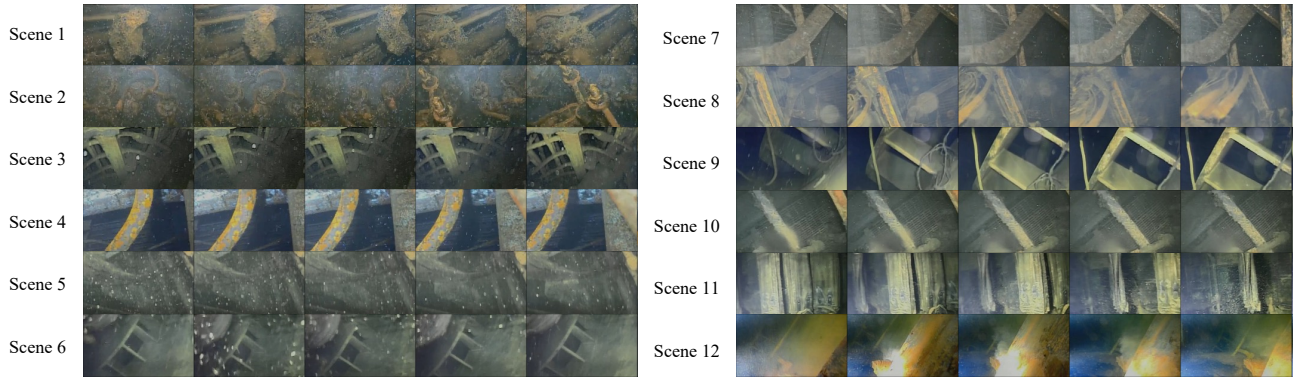


Figure 4. Example of video frame of survey videos of the PCV. (Source: TEPCO [16-19])

Fig.3(c), the result of the 3D reconstruction adequately reconstructed the structure with 19,461,828 points. The processing time for the 3D reconstruction was approximately 30,348s. Fig.3(d) shows the results of assigning the grouping labels to the reconstructed points. These snapshots were taken with the camera position roughly aligned with that of the image in Fig.3(a). The results show that the prediction results are reasonably assigned to the 3D points, showing that the proposed method works reasonably.

Here, we evaluated the correctness of the predicted group labels using the category labels of the 3D Computer-Aided Design model of the RTF mockup plant created by a specialized vendor. Based on the method of our previous study, a 3D point cloud of 94,952,392 points measured by FARO Focus S350 [14] was assigned nine labels (handrail, pipe, grating, equipment, fire suppression, light fitting, support, tank, and frame) as Ground Truth labels [6]. The category of equipment includes pump and bulb; frame include wall, door, floor, stair, checkered steel plate, and rusty structure; handrail include ladder. The scales of the Ground Truth 3D point cloud and the reconstructed 3D point cloud were roughly aligned manually and then aligned using Iterative Closest Point registration [15]. We evaluated the label of the reconstructed point as the closest on the K-Nearest Neighbor (KNN) algorithm by the labels of the Ground Truth point. The maximum distance of the KNN was limited to 50 mm. Accuracy was defined as the correctness rate where the predicted group label was the same as one of the corresponding Ground Truth labels by the following:

$$Accuracy = \frac{p_{ii}}{\sum_{j=0}^{K} p_{ij}} \quad (1)$$

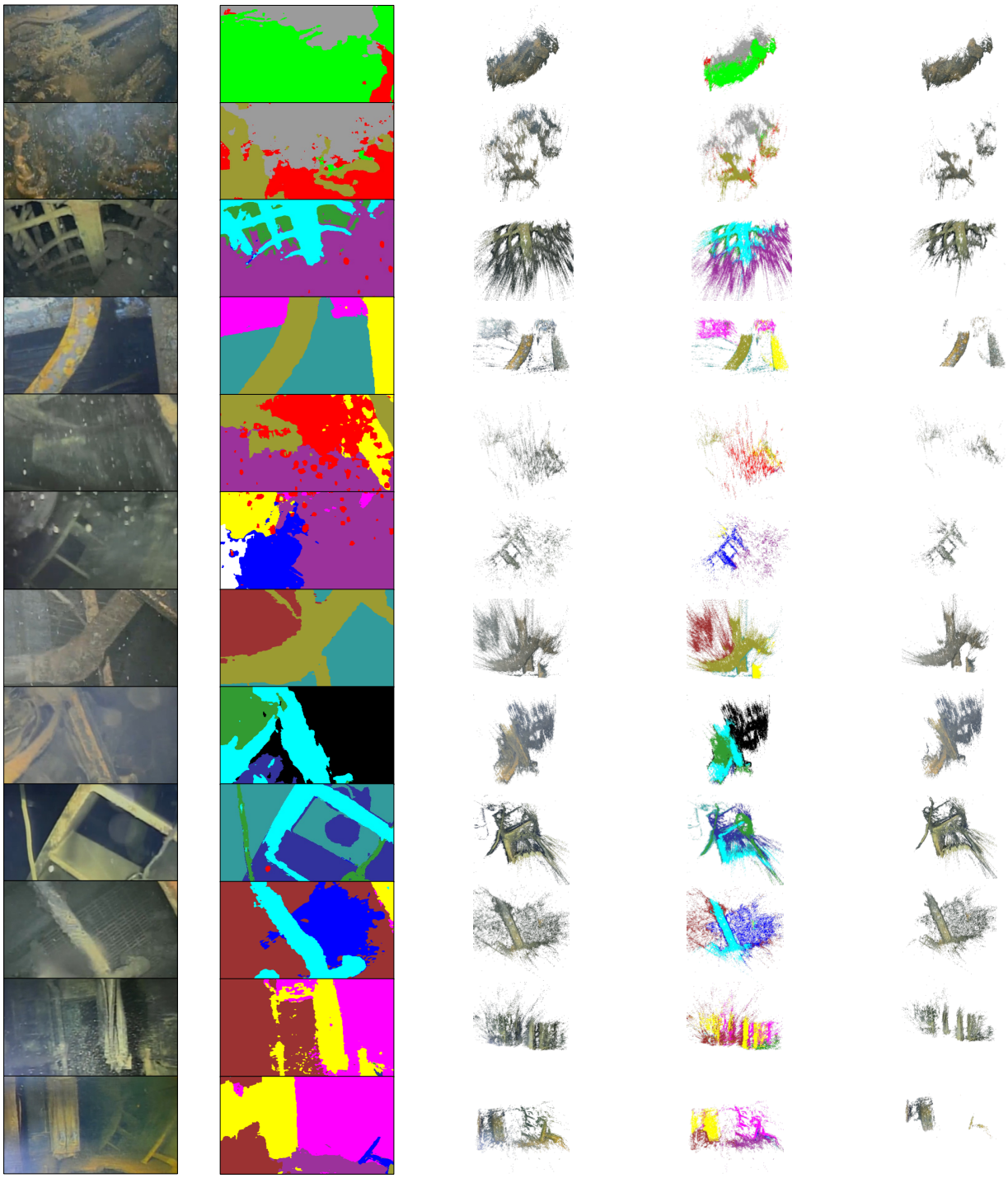
where K is the number of labels, and p_{ij} represents the point set for the prediction group label i, which belongs to the label j for the Ground Truth. Note that our purpose was not to evaluate the segmentation result but only to evaluate the correctness of the labels. The fourth column of Table 1 shows the calculated accuracy. The accuracy of most group labels was greater than 70%. The results show predicted group labels are close to the category separation that is generally used when human workers create 3D model data, and the results are similar to classification according to human attention. In some of the results, we found that changing in the lighting condition caused conflicts among labels from multiple camera angles. In particular, group labels 4, 8, and 10 contain many metallic appearances and are incorrectly predicted to be structures of other metallic appearances. In the future, we need to develop a method of label integration and a processing for changes in the lighting environment. The processing time, STEGO's training time was 5,071 s, prediction time was 3,258 s, and the time required for the labeling process in 3D reconstruction was 423 s, showing that our method can be introduced without significant time consumption.

TABLE II. RESULT OF PREDICTED GROUPS BY STEGO FOR THE PUBLISHED VIDEOS OF SURVEY IN THE PCV

No.	Color	Group contents
1		Floating object, radiation noise
2		Ceiling structures
3		Underwater structures (hazy by turbidity)
4		Pillar-like structures
5		Frame-like structures
6		Unclear region (turbid water, steam, etc.)
7		Unclear region (turbid water, steam, etc.)
8		Thin structures, dark place
9		Planer object, floating object near camera
10		Pipe-like structures
11		Unclear region (turbid water, steam, etc.)
12		Unclear region (turbid water, steam, etc.)
13		Grating-like structures
14		Unclear region (turbid water, steam, etc.)
15		Dark place

B. Result for Published Video of Survey in the PCV

Then, we applied 3D reconstruction based on grouping similar structure to survey videos of the PCV, which is the main objective of this study. The target was the video images of the survey inside the PCV of Unit 1 conducted in February, March, and May 2022 and March 2023, published by Tokyo Electric Power Company Holdings (TEPCO) [16-19]. These videos were captured by remote control of an underwater ROV. In this study, several scenes were selected and extracted from the published videos, taking into consideration various effect on images in the PCV. Fig.4 shows selected 12 short scenes. These scenes were captured by airborne and underwater cameras equipped on the ROV and were selected to be relatively moving ROV to generate 3D reconstruction model. Each scene is affected by radiation-affected imaging noise (scenes 1, 2, and 7), turbid water (scenes 6, 9, and 12), and floating objects (scenes 3, 5, 6, 8, and 11). We removed unnecessary caption of published video such as the capture date and time for reconstruction by trimming. The image size was 560 × 315 pixels. The STEGO model was trained using the same setup and computer as in the verification of typical plant environment. The total number of images used for training was 8,975.



(a) Input image (b) Grouping labels (c) 3D reconstruction result (d) 3D reconstruction result (grouping labels) (e) 3D reconstruction result (Only selected)

Figure 5. Result of grouping 3D reconstruction for survey videos of the PCV. (Source of (a): TEPCO [16-19])

Fig.5 shows the grouping and 3D reconstruction results. The input image (Fig.5(a)) was segmented into regions that appear to be structures and background regions, such as turbid water and grating in dark place (Fig.5(b)). In Fig.5(c) and 5(d), the results show that group labels were reasonably applied to the 3D reconstruction results. Table 2 shows the results of reviewing all images and confirming the objects contained in the 15 group. The shaded rows in Table 2 represent groups that the authors identified as structures, while the others represent groups that were difficult to identify structures. Each label was able to group regions of radiation noise and floating objects (group 1) and unclear regions (group 6, 7, 11, 12, and 14), which is generally in line with our aim. Fig.5(e) shows the result of displaying the reconstructed points using only the group labels that were determined to be structures. Compared to the point cloud displaying all group labels (Fig.5(c)), the contours of the structures are easier to recognize because the noise and artifacts generated between the structures are not displayed. The results of scenes 2 and 5 clearly show that floating objects and noisy regions can be hidden. These results

show that it is effective to segment the structural region and separate the 3D reconstruction results for our purpose. As an application, the operator can choose whether to show or hide group labels while comparing the predicted image of STEGO and the label of the 3D reconstruction result, which leads to improved visibility and a better understanding of the condition of the workspace.

On the other hand, the following issues need to be addressed in the future. First, at present, we have not yet identified the groups that generate noise in 3D reconstruction results, it is necessary to quantitatively evaluate accuracy. Also, we would like to introduce the human process of selecting the necessary group labels. However, because the perspectives of different humans are ambiguous, it is necessary to develop systems that integrate quantitative analysis and human judgment. In addition, further evaluation is required in a mockup environment under various conditions, such as the effects of noise and changes in lighting, to clarify the effectiveness and performance of the proposed method in a real environment. Next, we need to consider shortening the processing time. In this study, the processing time was increased simply by introducing the grouping module, although this was not remarkable. Considering the background of our study, it is easy to imagine that the computation time of 3D reconstruction would be very long when processing video images that take several hours in the survey. We consider that the computation time for 3D reconstruction can be reduced by using masks created by predicted images for groups that are judged to be unnecessary, thereby limiting the region to be used for processing, such as key-point extraction and triangulation. Finally, we need to analyze the optimal number of group labels because the number of group labels to be predicted is fixed for preliminary verification. In future work, we would like to report on the methodology and validation results of these ideas.

IV. Conclusion

In this study, we describe a 3D reconstruction method based on grouping similar structures for survey videos in the PCV of FDNPS. Our method groups uncertain structures in images by unsupervised learning, and the grouping results can be used to separate the 3D reconstruction results into groups. We verified that the proposed method works reasonably well by using video images of typical plant environments and the survey of the PCV. The 3D point clouds created by the proposed method can be utilized by an application in which human select the necessary labels, for example, improving the visibility of 3D reconstruction results in 3D viewers and eliminating noise for 3D modeling. In future, we need to develop an algorithm to identify groups that generate noise and verify the validity of the proposed method. We also plan to develop a method to reduce computation time by creating a mask from grouping labels and performing 3D reconstruction, and we will consider applying the proposed method to longer survey videos.

Acknowledgment

This work was supported by the JAEA Nuclear Energy Science & Technology and Human Resource Development Project Grant Number JPJA23P23813888.

References

[1] Mid-and-Long-Term Roadmap towards the Decommissioning of TEPCO's Fukushima Daiichi Nuclear Power Station. Japan: Tokyo Electric Power Company Holdings, Inc. https://www.meti.go.jp/english/earthquake/nuclear -/decommissioning/pdf/20191227_3.pdf

[2] T.Hanari and K.Kawabata: "3D Environment Reconstruction Based on Images Obtained by Reconnaissance Task in Fukushima Daiichi Nuclear Power Station," EJournal of Advanced Maintenance, Vol.11, No.2, pp. 99–105, (2019).

[3] H. C. Longuet-Higgins: "A computer algorithm for reconstructing a scene from two projections", Nature, Vol.293, No. 5828, pp.133-135 (1981).

[4] C. H. Esteban and F. Schmitt: "Silhouette and stereo fusion for 3D object modeling", Comput. Vis. Image Underst., Vol. 96, No.3, pp.367-392 (2004).

[5] P. Sarlin, C. Cadena, R.Y. Siegwart, and M. Dymczyk, "From Coarse to Fine: Robust Hierarchical Localization At large Scale," The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2019), (2018).

[6] T. Imabuchi, Kawabata, "Discrimination of the Structures in a Plant Facility Based on Projected Image Created from Colored 3D Point Cloud Data", The 2023 IEEE/SICE International Symposium on System Integration (SII 2023), 2023.

[7] M. Hamiltonet. et.al., Unsupervised Semantic Segmentation by Distilling Feature Correspondences, International Conference on Learning Representations (ICLR2022),2022.

[8] C. Mathilde, T. Hugo, M. Ishan, J. Herve, M. Julien, B. Piotr, J. Armand, "Emerging Properties in Self-Supervised Vision Transformers," Proceedings of the International Conference on Computer Vision (ICCV), (2021).

[9] K. Alexander, M. Eric, R. Nikhila, M. Hanzi, R. Chloe, G. Laura, X. Tete, W. Spencer, B.C. Alexander, L. Wan-Yen, D. Piotr, G. Ross, Segment Anything, arXiv:2304.02643, 2023.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich. "SuperPoint: Self-supervised interest point detection and description," In Workshop on Deep Learning for Visual SLAM, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2018).

[11] P.E. Sarlin and D. DeTone, T, Malisiewicz and A. Rabinovich, "SuperGlue: Learning Feature Matching with Graph Neural Networks," The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020), (2019).

[12] M. Ester, H.P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), pp.226–231, (1996).

[13] Infrastructure inspection and disaster response robot facilities, Facilities & Equipment Fees, Fukushima Robot Test Field, https://www.fipo.or.jp/robot/facility/infrastrucrure

[14] FARO 3D laser scanner, https://www.faro.com/ja-JP

[15] P.J. Besl, N.D. McKay, "Method for registration of 3D shapes," Proceedings of Sensor Fusion IV: Control Paradigms and Data Structures, Vol.1611, pp.586-606, (1992).

[16] Tokyo Electric Power Company Holdings: "Photo collection, Implementation Status of Unit 1 Primary Containment Vessel Internal Investigation (ROV-A2) (work conducted between May 19) at the Fukushima Daiichi Nuclear Power Station, https://photo.tepco.co.jp/en/date/2022-e/202205-e/220523-01e.html

[17] Tokyo Electric Power Company Holdings: "Photo collection, Implementation Status of Unit 1 Primary Containment Vessel Internal Investigation (ROV-A2) (work conducted between March 14~16) at the Fukushima Daiichi Nuclear Power Station, https://photo.tepco.co.jp/en/date/2022-e/202203-e/220324-01e.html

[18] Tokyo Electric Power Company Holdings: "Photo collection, Implementation Status of the Unit 1 Primary Containment Vessel Internal Investigation (As of February 10) at the Fukushima Daiichi Nuclear Power Station, https://photo.tepco.co.jp/en/date/2022-e/202202-e/220210-01e.html

[19] Tokyo Electric Power Company Holdings: "Photo collection, The Unit 1 Primary Containment Vessel (PCV) Internal Investigation (Using ROV-A2, Day 3, Part 2), https://photo.tepco.co.jp/en/date/2023-e/202304-e/230404-01e.html