Digital Life
The Conference of Digital Life vol.3

Proceedings Presentation

Room A
Chairperson:
- Mitsunori Makino (Chuo University)
- Jiang Wenyuan (Osaka Sangyo University)

A-1. A Study on the Comparison of Realism in Pedestrian Behavior Using LiDAR and Video Camera
Seigo Ouchi (Graduate School of Osaka Electro-Communication University) and Masaya Nakahara (Osaka Electro-Communication University)

A-2. Research for Extracting Activity Areas of Preschool Children in Early Childhood Education
Haruka Nishikoba (Kansai University Graduate School), Shigenori Tanaka (Kansai University), Tatsuma Iwamoto (Kansai University Graduate School), Yuhi Kameda (Kansai University), Eriko Wada (Kansai University) and Hiromi Taniguchi (Kansai University)

A-3. Basic Research on Estimating Sukiyaki Ingredients Using Deep Learning
Ibuki Ikeda, Kazuma Sakamoto, Yoshihiro Ueda (Komatsu University) and Tomoya Senda (Graduate School of Komatsu University)

A-4. Basic Research on Tree Species Identification Methods for Broad-Leaved Trees Using Deep Learning
Yuya Hirata, Kazuma Sakamoto, Yoshihiro Ueda (Komatsu University), Iori Iwata, Riku Kaiba (Graduate School of Komatsu University), Sinnosuke Miyashita (Komatsu University)

A-5. A Survey on Applicability of Similar Image Retrieval Technique for Low-Light Images
Shota Yamashita (Graduate School of Osaka Electro-Communication University), Masaya Nakahara (Osaka Electro-Communication University), Yoshinori Tsukada (Reitaku University) and Yoshimasa Umehara (Setsunan University)

A-6. A Survey on Feasibility of Tunnel Space Maintenance Management Using 3D Gaussian Splatting
Takumu Kuhara (Graduate School of Osaka Electro-Communication University) and Masaya Nakahara (Osaka Electro-Communication University)

A-7. Exploratory Study on Interpretation of Visual Information in Regional Disaster Management Plans Using Vision-Language Models
Ryuma Kawakubo (Graduate School of Hosei University), Kenji Nakamura (Osaka University of Economics), Kazuma Sakamoto (Komatsu University) and Ryuichi Imai (Hosei University)

Room B
Chairperson:
- Takehiko Yamaguchi (Suwa University of Science)
- Toshio Teraguchi (University of Marketing and Distribution Sciences)

B-1. Basic Research on Quantitative Analysis of Eight Stages of Shooting Using Skeletal Estimation Technology
Terumi Kakukawa, Kazuma Sakamoto, Yoshihiro Ueda (Komatsu University), Iori Iwata and Fuya Shibata (Graduate School of Komatsu University)

B-2. Research on effectiveness of SEO measures estimated from daily fluctuations in search rankings
Sho Okado, Masaya Nakahara (Osaka Electro-Communication University) and Kazuma Sakamoto (Komatsu University)

B-3. Research on relationship between categories of posted videos using LDA
Kyoya Takiguchi (Graduate School of Osaka Electro-Communication University), Masaya Nakahara (Osaka Electro-Communication University) and Sakamoto Kazuma (Komatsu University)

B-4. Text Mining on Benefits and Challenges of AI-Generated Academic Paper Short Videos and Text Summaries
Hayato Sezaki (Graduate School of The University of Tokyo), Takashi Goto (IBM Japan), Ayako Kurono (Fukunaga) (Graduate School of Iwate University), Hideo Kawamata (Graduate School of Kobe University) and Kayoko Kurita (The University of Tokyo)

B-5. Fundamental Study on Indoor Space Representation Using Local Spatial IDs Derived from Mobile Terminal Point Clouds
Ryo Komiya (Graduate School of Hosei University), Kenji Nakamura (Osaka University of Economics), Yoshinori Tsukada (Reitaku University), Yoshimasa Umehara (Setsunan University), Yasuhito Niina (Asia Air Survey Co., Ltd.) and Ryuichi Imai (Hosei University)

B-6. State changes when tick size is changed
Hiroyuki Maruyama (Takushoku University)

B-7. Basic Research on the Detection of Dust Mask Wearing Status
Wenyuan Jiang (Osaka Sangyo University), Yuhei Yamamoto (Kansai University), Hajime Tachibana, Keisuke Nakamoto, Kunihiro Katai (Komaihaltec Inc.), Daito Yosumi, Atsuki Yoshida and Aito Yoshikawa (Osaka Sangyo University)

Digital Life

Journal of Digital Life 2025, A-1

Conference Proceedings

A Study on the Comparison of Realism in Pedestrian Behavior Using LiDAR and Video Camera

Seigo Ouchi [1]and Masaya Nakahara [2]

[1] Graduate School of Information Science and Arts, Graduate School of Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan

1. Introduction

The COVID-19 pandemic has led to restrictions on in-person contact to prevent the spread of infection. Consequently, many events have been canceled, postponed, or limited in attendance. This has spurred a rapid expansion of online alternatives, with university open campuses being significantly impacted. During the pandemic, participation in conventional on-site open campuses declined, while those held online saw increased engagement. A key development among these alternatives was the digital campus: a representation of a university in a virtual space that allows users to explore and experience the campus remotely. In these digital campuses, real-world structures are typically reproduced by manually constructing them in a virtual reality space, a process that demands significant time and cost. To address this, previous research (Gao et al., 2025) has proposed a method for generating 3D models from images captured by Unmanned Aerial Vehicles (UAVs) and visualizing them on the web.

However, existing digital campuses may feature Non-Player Characters (NPCs), but they fail to replicate the authentic movements of students and faculty. This makes it difficult for prospective students to envision the daily atmosphere of campus life, such as the bustling environment of the cafeteria, which are important factors when choosing a university. Therefore, a technology is needed that not only reproduces the university's architecture but also includes scenes of daily use to enhance the realism of the digital campus. Existing research (Ouchi et al., 2024) has proposed methods for reproducing pedestrian flow using LiDAR (Light Detection and Ranging), considering the state of pedestrians. This method is reported to track pedestrians from point cloud data and reproduce their walking states and flow based on movement trajectories, thereby enhancing pedestrian realism. However, specific pedestrian behaviors such as limb movements and sitting actions have not been reproduced. As a result, it is not possible to experience the typical liveliness and realism in areas where students tend to congregate, such as cafeterias and lobbies.

2. Methods

To address the limitations of existing methods, this study employs deep learning to estimate the skeletal motion of pedestrians from video footage and reproduces their behavior within a digital campus. Specifically, we use PromptHMR (Promptable Human Mesh Recovery) (Yufu et al., 2025) to reconstruct pedestrian behavior from captured footage. PromptHMR is a deep-learning-based technology that estimates a pedestrian's 3D skeletal motion from a single-view video by analyzing the spatial relationship between the camera and the person. The motion data from the 3D model output by PrompHMR is then applied to a pre-prepared 3D avatar model. Finally, each avatar is manually positioned and scaled within the 3D building data, as illustrated in Fig. 1.

3. Demonstration experiment

In this experiment, we conducted a questionnaire survey with 15 university students. They were shown three types of videos: one featuring pedestrian behavior generated by our proposed method, one with pedestrian flow from an existing method (Ouchi et al., 2024), and a statistical video. The survey aimed to verify the effectiveness of the proposed method.

Published: 6 September 2025
* Correspondence: nakahara@oecu.jp

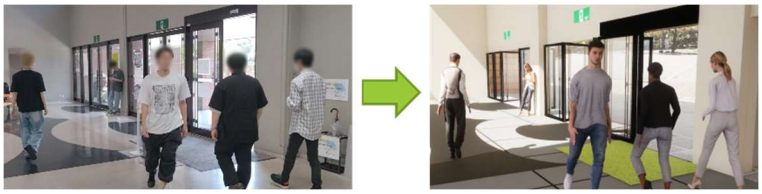


Fig. 1. Visualization results of behavior of pedestrians in virtual reality space

Table 1. Survey results on realism (number of people)

Video type	Feels unnatural	Feels natural	Feels realistic	Feels unrealistic
Statistical video	9	4	2	13
Existing method	4	4	2	2
Proposed method	2	7	11	1

Table 2. Results of the proposal method survey (number of people)

Survey contents	Strongly disagree	Disagree	Agree	Strongly agree
The avatar's limb movement feels unnatural	6	3	4	2
The movement behavior feels unnatural	5	7	2	1
The movement looks like that of a real person	1	0	8	6
It feels realistic	1	2	5	7

The survey results regarding realism are shown in Table 1. The results indicate that the proposed method was rated more highly than the other videos. Eleven participants rated it as "realistic," while only two found it "unnatural." This confirms the effectiveness of the proposed method. The survey results for the proposed method are detailed in Table 2. These results suggest that the proposed method can effectively reproduce everyday scenes. For instance, a total of 14 participants "agreed" or "strongly agreed" that the movement resembled a real person and that the scene felt realistic. However, some movements were perceived as unnatural. This is likely due to discrepancies between the avatar's height and the actual pedestrian's height, as all avatars were of a uniform size.

4. Conclusions

In this study, we conducted a comparative survey on the realism of pedestrian behavior generated by an existing LiDAR-based method (Ouchi et al., 2024) and our proposed video-based method. The empirical experiment confirmed the effectiveness of our proposed method. Currently, the position and scale of the 3D pedestrian behaviors are manually adjusted to align with the 3D building model. However, performing this manual adjustment for an entire campus is a significant challenge. Therefore, future work includes developing a method to automatically align the position and scale with the 3D model based on features extracted from the video. Furthermore, as creating 3D building models is time-consuming, we plan to integrate our method with advanced techniques for constructing highly realistic 3D spaces, such as 3D Gaussian Splatting (Bernhard et al., 2023). This integration aims to streamline the creation process and facilitate the construction of a more realistic digital campus. By addressing these issues, we aim to further improve the proposed method and enhance its realism.

References

Bernhard, K., Georgios, K., Thomas, L., George D. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering, *ACM Transactions on Graphics (TOG)*, *42*(4), pp.1-14.

Gao, R., Yan, G., Wang, Y., Yan, T., Niu, R., Tang, TC. (2025). Construction of a Real-Scene 3D Digital Campus Using a Multi-Source Data Fusion. A Case Study of Lanzhou Jiaotong University, *ISPRS Int. J. Geo-Inf, 14*(1).

Ouchi, S., Nakahara, M. (2024). A Study on the Sense of Presence of a Digital Campus Using LiDAR Considering Walking Conditions, *Annual Conference of the Virtual Reality Society of Japan, 29*.

Yufu, W., Yu, S., Priyanka, P., Kostas D., Michael, J. B. and Muhammed, K. (2025). Promptable Human Mesh Recovery. *arXiv*. https://arxiv.org/abs/2504.06397v2.

Digital Life

Journal of Digital Life 2025, A-2

Conference Proceedings

Research for Extracting Activity Areas of Preschool Children in Early Childhood Education

Haruka Nishikoba [1], Shigenori Tanaka [2], Tatsuma Iwamoto [1], Yuhi Kameda [2], Eriko Wada [2] and Hiromi Taniguchi [2]

[1] Graduate School of Informatics, Kansai University Graduate School, 2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka 569-1095, Japan
[2] Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka 569-1095, Japan

1. Introduction

In Japan, although the number of children using childcare facilities is decreasing due to the declining birthrate and aging population, the number of such facilities continues to increase. Meanwhile, the number of employees engaged in early childhood education has been declining year by year, making the reduction of their workload a pressing issue. Although, the Children and Families Agency revised staffing standards for each age group in 2024, the number of children assigned to each staff member remains high, and further improvement in the "Quality of education" is required. As a potential solution, the utilization of ICT in early childhood education is drawing attention for alleviating the workload. Nevertheless, current applications of ICT are mostly limited to child education using tablets and administrative support for paperless operations. The implementation of ICT in tasks involving direct contact with children such as child monitoring has not yet been realized. Although some facilities have implemented live video streaming, this approach has introduced new tasks such as monitoring and analyzing the footage and as a result has not effectively reduced the workload. This research aims to support the monitoring of preschool children in early childhood education facilities by developing technologies for child detection and class classification using AI and camera images installed indoors. Furthermore, to present the acquired data in an intuitive and comprehensible manner, we explore methods for visualizing the data within a metaverse environment, ultimately aiming to develop technology that is practical and applicable in real world settings.

2. Proposed method

The processing flow of the proposed system is illustrated in Figure 1. In this research, we utilize two days of video footage captured by two cameras installed in a kindergarten (Figure 2). First, we construct a detection model that classifies two categories (children and adults) by fine-tuning the existing YOLO v7 model with our original dataset.

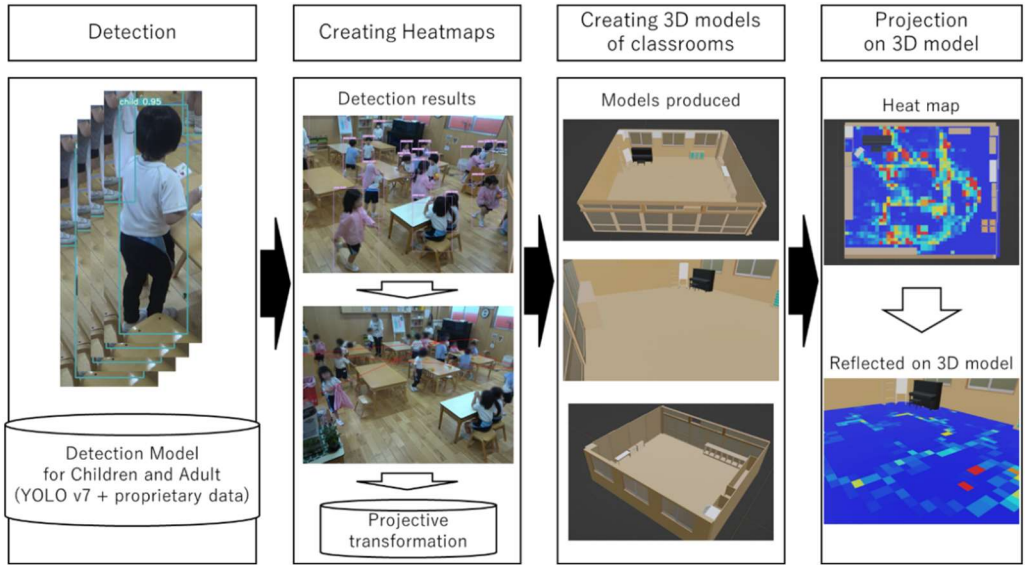


Fig. 1. Process flow of developed system

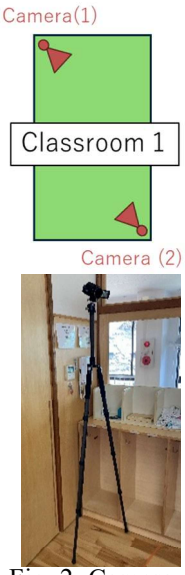


Fig. 2. Camera

Published: 6 September 2025
* Correspondence: k191775@kansai-u.ac.jp;

Next, a perspective transformation is applied to the captured images. Furthermore, a 3D model of the classroom is created using point cloud data captured on a different day at the same facility. Finally, we generate a pseudo heatmap by altering the floor colors on the subdivided 3D model based on the transformed detection data. This visual representation enables an intuitive understanding of the detection results.

3. Demonstration experiment

3.1 Experimental setup

This experiment evaluates the effectiveness of the proposed method using footage from Camera (2) in Figure 2. A custom trained model with YOLO v7 is used to detect and classify children and adult. The detection results are manually verified. After applying a perspective transformation, the results are projected onto a 3D model of the classroom.

3.2 Results and Discussion

Figure 3 shows the evaluation of detection accuracy. Overall, the estimated values tended to be lower than the ground truth, likely due to missed detections from occlusions by adults or lockers and children being partially out of frame. This issue could be mitigated by improved camera placement or using multiple cameras.

Several false positives were also observed where adults were misclassified as children (Figure 4), likely due to the smaller amount of adult data. Adding more such data and incorporating temporal information such as tracking are expected to further improve accuracy. Figure 5 also shows the cumulative detection results visualized as a heatmap on the 3D model. This technique could be used to identify children's activity and activity areas in classrooms, supporting the optimization of furniture and equipment layouts and revealing class specific behavioral patterns through long term data accumulation.

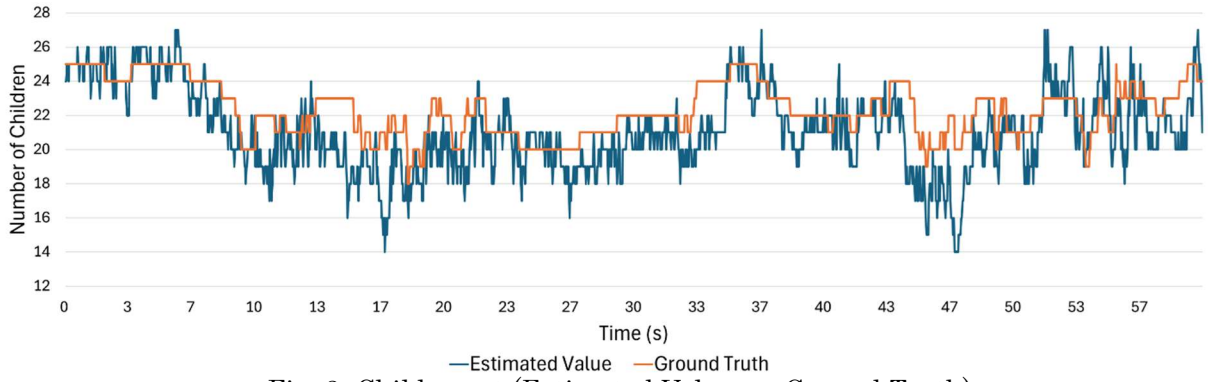


Fig. 3. Child count (Estimated Value vs. Ground Truth)

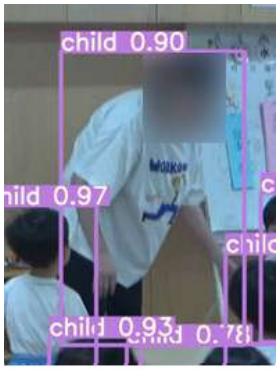


Fig. 4. Examples of false positives

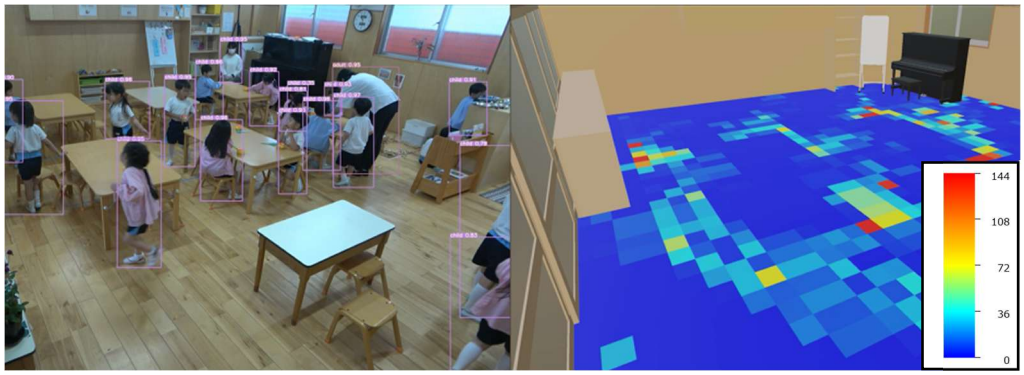


Fig. 5. Cumulative heatmap

4. Conclusion

In this research, we developed the method to detect children and adults from camera footage in early childhood education settings and visualize the results in a 3D space. The custom trained model enabled the detection of children's positions, and the results could be visualized as the pseudo heatmap within accurately the 3D environment. This visualization allows for intuitive understanding of individuals' locations in the classroom and has the potential to support staff in child monitoring tasks. In future works, we plan to improve the detection accuracy for staff members and incorporate person tracking techniques to evaluate the applicability of the system to continuous video streams.

Digital Life

Journal of Digital Life 2025, A-3

Conference Proceedings

Basic Research on Estimating Sukiyaki Ingredients Using Deep Learning

Ibuki Ikeda [1], Kazuma Sakamoto [1*], Yoshihiro Ueda [1] and Tomoya Senda [2]

[1] Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[2] Graduate School of Sustainable Systems Science, Komatsu University, Komatsu, Ishikawa 923-8511, Japan

1. Introduction

Computer vision has advanced for visual data analysis with applications in healthcare, manufacturing, and autonomous driving. In food analysis, it enables calorie estimation by identifying food categories and portion sizes. Im2Calories (Myers et al., 2015) introduced CNN-based depth and volume estimation from a single image, and later studies adopted multi-task frameworks integrating recognition, segmentation, and volume estimation. In recent years, stereo cameras have achieved higher precision and have been applied to food volume measurement. However, research utilizing high-precision stereo cameras has so far been limited to food volume estimation, and studies focusing on ingredient-level volume measurement and calorie estimation have not been conducted. A key challenge in such measurement is the necessity of manual region specification. In this research, we aim to estimate calories by detecting ingredients with YOLOv8n-seg, reconstructing point clouds using a stereo camera to calculate their volumes, and then performing calorie estimation. It is considered that high-precision calorie estimation can be achieved by performing accurate volume measurement at the ingredient level, enabling category-independent estimation that accounts for differences among individual ingredients. Specifically, we estimated five representative sukiyaki ingredients and the sukiyaki pot region using YOLOv8n-seg, and reconstructed point clouds with the ZED2i stereo camera.

2. Methods

Fig. 1 shows an overview of this research. Solid lines indicate evaluated components, and dashed lines represent unimplemented ones. Using sukiyaki images as input, ingredient and sukiyaki pot region estimation was performed with the Ingredients Estimation Model and the Sukiyaki Pot Estimation Model, respectively, while point clouds were generated through Point Cloud Measurement Using the ZED2i Stereo Camera. The ultimate objective is to estimate calories with the Calorie Estimation Model, integrating the estimation and measurement results from these three models.

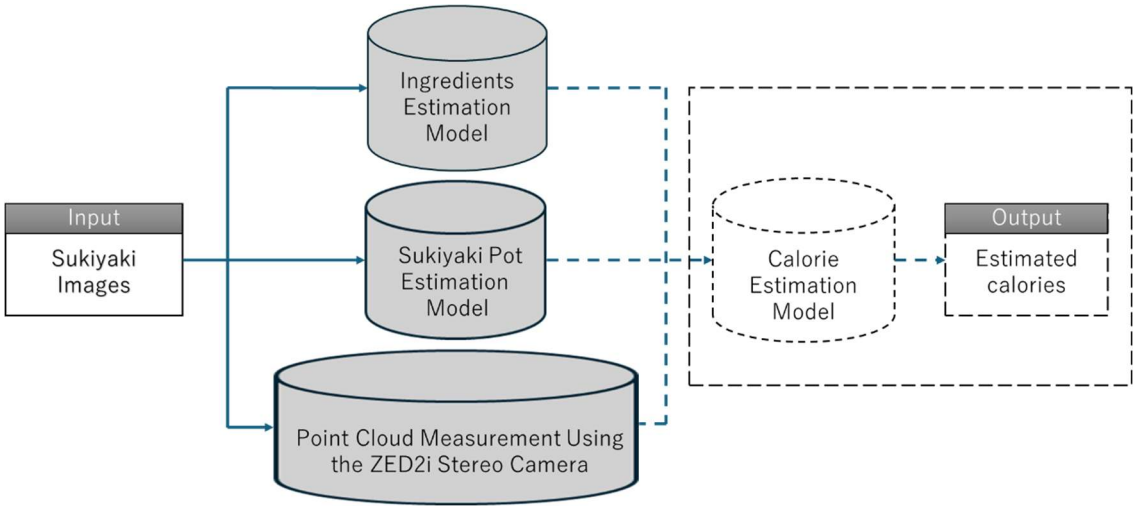


Fig.1. Overview of this research

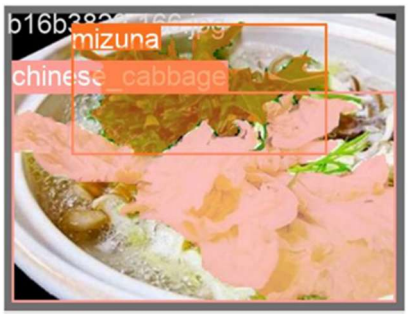
Published: 6 September 2025
* Correspondence: kazuma.sakamoto@komatsu-u.ac.jp;

Sukiyaki images were used for ingredients and sukiyaki pot estimation with YOLOv8n-seg, while point clouds were generated using the ZED2i stereo camera. For ingredients estimation, five typical sukiyaki ingredients (beef, Chinese cabbage, mizuna, shiitake, and shirataki) were targeted. The dataset comprised 734 sukiyaki and 500 hotpot images, totaling 1,234 images for five-fold cross-validation. Training used 200 epochs, batch size 8, with EarlyStopping after 100 stagnant epochs. For sukiyaki pot estimation, the whole pot was the target, with 266 images collected and five-fold cross-validation conducted under consistent conditions. Point cloud measurement was performed with the ZED2i placed 30 cm above the table. A total of 25 point clouds were obtained, coordinates in millimeters, and pot diameter and height were calculated for evaluation using mean absolute error (MAE).

3. Results and Discussion

We first present the Results and Discussion of ingredients estimation. The overall mAP@0.5 across all classes was 0.762. Mizuna and shiitake achieved relatively high scores, with mAP@0.5 values of 0.840 and 0.874, respectively, whereas beef, Chinese cabbage, and shirataki obtained lower scores of 0.742, 0.668, and 0.687, respectively, indicating a performance gap of approximately 0.10. The likely reason mizuna and shiitake achieved higher scores is that mizuna requires minimal cooking, and shiitake retain their appearance even after heating, resulting in many images with consistent visual appearances. In contrast, beef, Chinese cabbage, and shirataki undergo significant appearance variations due to heating, which likely contributed to their lower scores. Fig.2 illustrates an example in which Chinese cabbage was not detected, supporting the interpretation that large appearance variations prevented the model from responding effectively. As a potential approach to address this issue, incorporating training data that capture various degrees of cooking could be considered. Next, we present the Results and Discussion of sukiyaki pot estimation. Sukiyaki pot estimation achieved mAP@0.5 of 0.995. The result indicates that sukiyaki pot regions were estimated with high accuracy. Fig.3 presents an example of the sukiyaki point cloud measured with the ZED2i. The MAE of the point cloud measurements was 0.80 mm for diameter and 1.28 mm for height, indicating that the actual sukiyaki could be reproduced as a point cloud with high accuracy. The point clouds obtained with the ZED2i demonstrated high precision. In addition, sukiyaki pot estimation can be performed with high accuracy under the condition that the same pot is used. Therefore, by removing regions outside the sukiyaki pot from the acquired point cloud, it is considered feasible to construct a point cloud limited to the sukiyaki pot region. Although ingredients estimation faces challenges in handling appearance variations caused by different degrees of cooking, we consider that incorporating data representing diverse cooking conditions could improve the accuracy of ingredients estimation, thereby enabling the realization of a system that estimates calories based on the volume of each ingredient.

.



(a) Ground-truth Label

(b) Predicted Label

Fig.2. Example in which Chinese cabbage was not detected

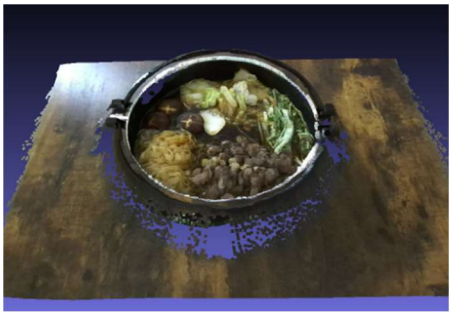


Fig.3. Example of the sukiyaki point cloud measured with the ZED2i

4. Conclusions

In this research, ingredients estimation and sukiyaki pot estimation were conducted using YOLOv8n-seg with five-fold cross-validation. Sukiyaki was also captured with the ZED2i stereo camera to obtain point clouds, from which the pot's diameter and height were measured. While ingredients estimation still presents challenges, sukiyaki pot estimation achieved near-perfect accuracy, and the ZED2i point clouds were shown to be highly reliable. By combining region estimation with YOLOv8n-seg and point clouds from the ZED2i, we consider it feasible to develop a calorie estimation system. Future work will focus on completing and evaluating this integrated system.

References

Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., & Murphy, K. (2015). Im2Calories: Towards an automated mobile vision food diary. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1233–1241. https://doi.org/10.1109/ICCV.2015.146

Digital Life

Journal of Digital Life 2025, A-4

Conference Proceedings

Basic Research on Tree Species Identification Methods for Broad-Leaved Trees Using Deep Learning

Yuya Hirata [1], Kazuma Sakamoto [1]*, Yoshihiro Ueda[1],Iori Iwata[2], Riku Kaiba[2], Sinnosuke Miyashita[3]

[1]Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[2]Graduate School of Sustainable Systems Science, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[3]Formerly Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa 923-8511, Japan

1. Introduction

Modern Japanese forestry faces challenges such as the decline in coniferous timber prices and a decrease in the workforce. As a result, forestry using Broad-Leaved Trees has attracted attention; however, fundamental data on their resource volume and distribution remain insufficient. Conventional surveys require considerable time and labor for manual Tree Species Classification, making continuous implementation difficult. Regarding Tree Species Classification, (Koubara et al., 2021). proposed a method using CNNs to classify bark images of coniferous species such as Cryptomeria japonica and Chamaecyparis obtusa, achieving an identification rate of about 80%. Nevertheless, research focusing on Broad-Leaved Trees remains insufficient. Therefore, this research applies image recognition to images of Broad-Leaved Trees captured by Drones. In doing so, it attempts remote Tree Species Classification and aims to improve the efficiency of forest surveys.

2. Methods

The proposed method takes as input forest videos captured by a Drone. The input videos are divided into frames, and inference is performed for each image using the constructed tree species classification model. The tree species classification model in this research was trained with YOLOv8 using images of trees recorded by a video camera. The drone used was DJI Avata 2(DJI, 2025), and the video camera was SONY FDR-AX60, which recorded in 4K resolution.In the experiments, a model was trained to classify bark images into three categories: "Magnolia obovata," "Quercus serrata," and "Others."

In experiment 1, the agreement rate between manually annotated data and the model's inference results was output as the training result, and accuracy evaluation and discussion were conducted. The evaluation metrics were precision, recall, and AP50.

In experiment 2, based on the results of experiment 1, the training data were augmented, and training was conducted again. For "Magnolia obovata" and "Quercus serrata," brightness values were modified for some of the bark images with shadows. For "Others," brightness values were modified for all images. The number of "Magnolia obovata" images was increased from 437 to 537 by applying multipliers of 1.5 and 0.5. The number of "Quercus serrata" images was increased from 703 to 768 by applying a multiplier of 0.5. The number of "Others" images was increased from 64 to 640 by applying multipliers of 1.5, 1.25, 0.75, and 0.5. The number of training epochs was set to 150, and the batch size was set to 24.

3. Results

The experimental results are shown in Table 1. Table 1 summarizes the values of precision, recall, and AP50 on the test data. As shown in Table 1(a), the results of experiment 1 indicated that overall precision was 80%, recall was 70%, and AP50 was 80%. However, the recall of the "Others" class was as low as 30%.

Published: 6 September 2025
* Correspondence: kazuma.sakamoto@komatsu-u.ac.jp;

As shown in Table 1(b), the results of experiment 2 indicate that, compared with experiment 1, the overall accuracy improved; however, the recall of "Magnolia obovata" and "Quercus serrata" decreased.

Table1. Accuracy of Identification for Tree Species

(a)Results of Experiment1

Class	Precision	Recall	AP50
All	0.801	0.718	0.812
Magnolia obovata	0.604	0.947	0.788
Querrcus serrata	0.830	0.894	0.910
Others	1.000	0.313	0.740

(b) Results of Experiment2

Class	Precision	Recall	AP50
All	0.860	0.814	0.891
Magnolia obovata	0.674	0.789	0.809
Querrcus serrata	0.907	0.886	0.937
Others	1.000	0.766	0.926

4. Discussion

The results of experiment 1 indicated that the recall of the "Others" class was notably low. This phenomenon is believed to be attributable to inadequate learning, stemming from the limited amount of training data available. It is expected that improvement can be achieved by applying data augmentation in the future.

As a result of experiment 2, the AP50 value increased, indicating that data augmentation led to an overall improvement in accuracy. The precision of "Magnolia obovata" and "Quercus serrata" increased, showing that false detections were reduced. However, recall decreased. This phenomenon is attributed to the diminution of brightness values, a factor that has been demonstrated to engender a more stringent evaluation process and an augmented number of false negatives in target detection. This suggests that in this experiment, the brightness values were excessively reduced. In the future, it will be necessary to adjust the detection threshold and the balance of augmentation. The accuracy of the "Others" category demonstrated a substantial enhancement in comparison with experiment 1. This enhancement can be ascribed to the augmented quantity of training data samples in experiment 2, a factor that had considerably contributed to the accuracy decline of the "Others" category in experiment 1.

5. Conclusions

In this research, we proposed a method for tree species classification using deep learning. In our experiments, the setting of training data for the "Others" class posed challenges, making it difficult to achieve practical classification accuracy. Data augmentation is a practical approach, as it allows for increasing the amount of training data without the need for additional data collection if the original dataset exists. However, if the augmentation is not appropriate for the intended purpose, its effectiveness is limited; therefore, the number and type of augmented images must be carefully considered. In this research, we plan to explore effective augmentation methods, such as brightness adjustment and horizontal flipping. Based on the conclusions of this research, future work will focus on constructing training data from drone-captured images, examining classification methods for the "Others" class, and implementing effective data augmentation strategies.

Acknowledgments

We would like to express our sincere gratitude to the members of the NPO Green Web Ishikawa for their significant contributions to the progress of this research, to the members of the Kaga Forestry Cooperative for their cooperation in obtaining permission from the forest owners for data collection, and to Mr. Yutaka Yada of the Ishikawa Prefectural Forestry Experiment Station for his support during field data collection.

References: (APA Style)

DJI. (2025). DJI Avata2. Dji. https://www.dji.com/jp/avata-2

Forestry Agency of Japan. (2025). Trends in timber prices in Japan. Forestry Agency. https://www.rinya.maff.go.jp/j/kikaku/hakusyo/r1hakusyo_h/all/chap3_1_3.html

Koubara, A. Tominaga, A., Shigaki, S., Hayashi,E. Fujisawa, R. (2021). Classification of *sugi* and *hinoki* from bark images using deep learning with CNN. *Journal of Forest Utilization*, *36(1)*, 5–12. https://doi.org/10.18945/jjfes.36.5

Ultralytics. (2025). YOLO v8. Ultralytics. https://www.ultralytics.com/ja

Digital Life

Journal of Digital Life 2025, A-5

Conference Proceedings

A Survey on Applicability of Similar Image Retrieval Technique for Low-Light Images

Shota Yamashita [1], Masaya Nakahara [2], Yoshinori Tsukada [3] and Yoshimasa Umehara [4]

[1] Graduate School of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[3] Faculty of Engineering, Reitaku University, 2-2-1 Hikarigaoka, Kashiwa-shi, Chiba 277-8686, Japan
[4] Faculty of Business Administration, Setsunan University, 17-8 Ikedanakamachi, Neyagawa-shi, Osaka 572-8508, Japan

1. Introduction

A declining birthrate and an aging population have led to a labor shortage in the security industry, which has become a significant social issue. To address this issue, security systems utilizing surveillance cameras and other monitoring devices have become commonplace. However, when comprehensively monitoring complex facilities such as factories with numerous pieces of equipment, blind spots often occur due to occlusion, necessitating the installation of many cameras. Consequently, there is significant interest in automated patrols using drones. Previous research (Nakahara, M., et al., 2025) developed a method to estimate a drone's position by searching a pre-collected image database. The search is based on changes in the distribution of features and color information within images captured by an RGB camera. However, during nighttime hours when security is most critical, obtaining distinctive feature points from images is difficult, which complicates self-position estimation. Therefore, this study investigates the feasibility of position estimation via similar image search by applying a deep learning-based brightness correction method to nighttime images to extract color distribution information.

An existing method for autonomous nighttime navigation uses deep learning to predict the direction of travel from captured images (Kothari, P., et al., 2021), while another predicts the travel path using data from both thermal and RGB cameras (Aditya, NG., et al., 2024). This latter method uses a thermal camera, enabling autonomous nighttime navigation with accuracy comparable to that of an RGB camera during the day. However, data from thermal cameras has low resolution and cannot capture personal attributes like clothing, making it difficult to integrate with other security systems. Therefore, this study investigates using low-light images captured by an RGB camera with a portable light—a setup mountable on a small drone—for similar image retrieval and verifies the applicability of this technology for security drones.

2. Methods

This section describes the procedure for calculating image similarity using nighttime images as input, adapting a method from previous research. This study employs the dHash algorithm for similar image retrieval. dHash generates a binary hash value by dividing an image into regular intervals and capturing the color difference between adjacent regions. When applied directly to low-light images, it is expected that subtle changes in color information will not be captured, leading to failures in the similarity search. Therefore, this study investigates the effectiveness of two brightness correction approaches for enabling similar image searches at night: a deep learning-based method and a rule-based method.

The deep learning approach utilizes RUAS (Risheng, L., et al., 2021) to correct low-light images. However, nighttime images are often affected by shot noise. Therefore, after correction with RUAS, noise reduction is performed using an Unsharp Mask and a Gaussian Filter to prevent the generation of incorrect hash values. The rule-based method employs a classic technique: first, the overall image brightness is increased via gamma correction and then Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance contrast in dark regions. Subsequently, similar to the deep learning method, noise reduction is applied to the corrected image to prevent incorrect hash value calculations.

Published: 6 September 2025
* Correspondence: nakahara@oecu.jp

3. Demonstration experiment

In this experiment, the proposed methods were applied to images captured in a hallway at night. The results were then compared with a baseline where no brightness correction was used. The experiment aimed to verify if the corrected images could be used for position estimation for automated patrols by small drones equipped with RGB cameras. The experimental site was selected based on the assumption that the method would be combined with previous research methods that estimate self-position from images captured by an RGB camera, and a hallway within a university campus with many plain walls was chosen. In addition, a group of images to be used as search sources was collected prior to the start of the experiment, and similar images were searched for from the images captured during the experiment. For this procedure, three sets of results were confirmed: 1) without brightness correction, 2) with deep learning-based correction, and 3) with rule-based correction. The accuracy rate, defined as the percentage of correctly identified similar images from the input set, was calculated to evaluate the applicability of the proposed methods.

Fig.1 Examples of brightness correction results using each method

Table 1 Calculation results for the accuracy rate of each method

Method	Number of input images	Number of correct images	accuracy rate
Low-light image	60	39	65.0%
Deep learning	60	48	80.0%
Rule-base	60	46	76.7%

Fig. 1 shows an example of an image corrected using each method, and Table 1 lists the calculated accuracy rates. The results in Table 1 confirm that brightness correction improves the accuracy rate. In particular, the deep learning method achieved an 80% accuracy rate, confirming the feasibility of performing similar image searches at night using an RGB camera with a light source. This suggests that even low-illumination images can be utilized for automated drone patrols after brightness correction. However, we found that similar searches were difficult in certain sections. For instance, the search was successful for the scene in Fig. 1(a) but failed for the one in Fig. 1(b). This is because dHash, which is used for searching similar images, captures differences in color information across the entire image. Consequently, when the distance to the forward-facing wall is large, as in case (b), the subtle differences between consecutive images are difficult for the algorithm to capture, leading to unsuccessful searches. Therefore, future improvements will require incorporating methods sensitive to local changes, such as feature point-based approaches, to enhance search accuracy.

4. Conclusions

This study investigated the applicability of using brightness-corrected low-light images for similar image retrieval, aiming to enable indoor, nighttime security patrols by small drones with RGB cameras. Through empirical experiments, we demonstrated that the deep learning-based technique achieves an 80% accuracy rate for similar image retrieval. Future work will focus on improving search accuracy by combining the current approach with methods capable of responding to local image features, with the goal of practical application in security drones.

References

Aditya, NG., Dhruval PB., Jehan S., Shubhankar J., Xueji W., & Zubin J. (2024). Thermal Voyager: A Comparative Study of RGB and Thermal Cameras for Night-Time Autonomous Navigation, *ICRA*, 14116-14122.

Kothari, P., Kreiss, S., & Alahi, A. (2021). Human Trajectory Forecasting in Crowds: A Deep Learning Perspective, *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 7386-7400.

Nakahara, M., Tsukada, Y., Umehara, Y., & Yamashita, S. (2025). Research on Indoor Self-Location Estimation Technique Using Similar Image Retrieval Considering Environmental Changes, *Journal of Digital Life*, *5*(SpecialIssue).

Risheng, L., Long, M., Jiaao, Z., Xin, F., & Zhongxuan, L. (2021). Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement, *CVPR*, 10561-10570.

Digital Life

Journal of Digital Life 2025, A-6

Conference Proceedings

A Survey on Feasibility of Tunnel Space Maintenance Management Using 3D Gaussian Splatting

Takumu Kuhara [1] and Masaya Nakahara [2]

[1] Graduate School of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan

1. Introduction

Much of the social infrastructure in Japan, constructed during the period of rapid economic growth, is now deteriorating due to age. Consequently, the adoption of efficient maintenance and management technologies has become a pressing issue. For tunnels in particular, ensuring structural safety requires inspecting the walls for abnormalities such as cracks. Traditionally, laser scanners have been used to measure 3D point cloud data, capturing the spatial geometry inside tunnels. However, this method requires expensive equipment and specialized expertise, posing significant challenges in terms of cost and practicality.

Against this backdrop, 3D Gaussian Splatting (3DGS), a technique for reconstructing 3D models from images, has recently gained attention (Kerbl B., et al., 2023). This method can generate visually realistic 3D reconstructions. However, its applicability to environments with few distinctive features and minimal geometric variation, such as tunnels, has not been thoroughly investigated. Previous research by Kagaya, Y., et al. (2024) involved reconstructing and evaluating a 3D model of an approximately 25-meter section of a tunnel under construction. However, there is a lack of research on its application to long, operational tunnels where unrestricted photography is challenging. Therefore, this study investigates the feasibility of using 3DGS to reconstruct 3D models of tunnel spaces for maintenance and management applications. This study also examines the challenges of reconstructing tunnel spaces using 3DGS and explores potential solutions.

2. Methods

This section outlines the process used to reconstruct a 3D model of tunnel space using 3DGS. First, video footage of the entire tunnel was captured using a 360-degree camera while walking along the interior sidewalk. During filming, slow walking is used to minimize video shake caused by hand movement. In addition, to prevent the photographer from appearing in the split screen, the 360-degree camera is always held above head height during filming. The captured video was then processed to extract distortion-corrected images for various viewing directions (Fig. 1). These multiple images were then used as input to reconstruct the 3D model of the tunnel using 3DGS.

Fig. 1. Images of the tunnel divided and distortion-corrected from the 360-degree video

Published: 6 September 2025
* Correspondence: nakahara@oecu.jp

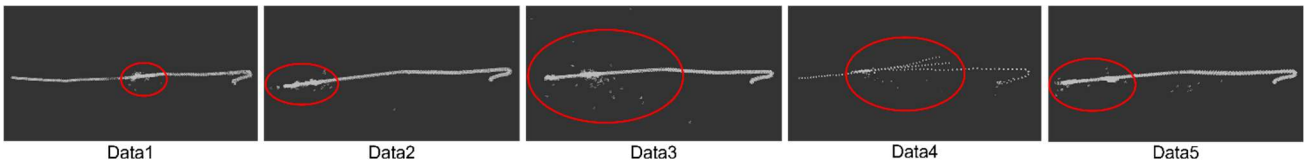


Fig. 2. Estimated shooting position result

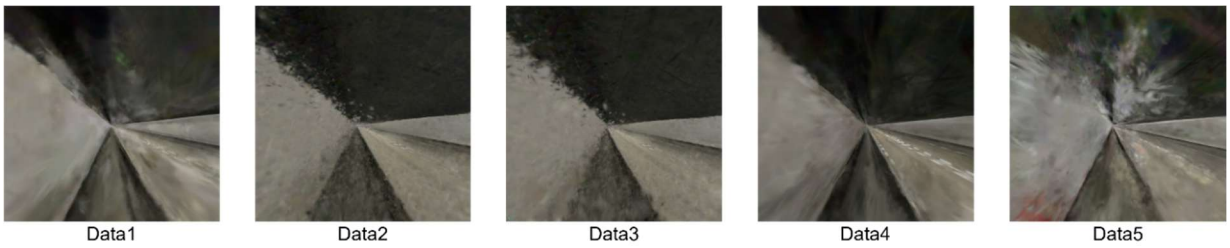


Fig. 3. Visual reconstruction results

3. Demonstration experiment

In our experiment, the 360-degree video was segmented into five different image sets (datasets), each of which was used to reconstruct a 3D model with 3DGS. The datasets were created as follows: Data1: Images from an eight-way horizontal split. Data2: Images from a twelve-way horizontal split. Data3: A combination of Data1 and Data2. Data4: Images from a nine-way split focused on the forward direction. Data5: A combination of images from a nine-way forward split and a nine-way backward split. For comparison, these same datasets were also processed using Agisoft Metashape, a commercial photogrammetry software package, to generate 3D spatial data. This setup allowed for a comparative analysis of the reconstruction capabilities of Metashape and 3DGS. The target tunnel for this study is mostly straight and approximately 1.1 km long. We recorded video footage along the walkway spanning the entire 1.1 km length of the tunnel. The estimated camera poses for each dataset are shown on Fig. 2. Red circles indicate regions where pose estimation failed. Data1: Failed near the center of the tunnel. Data2: Failed near the ends of the tunnel. Data3: Failed across a wide area. Data4: Failed significantly, resulting in multiple disjointed camera paths. Data5: Failed near one end of the tunnel. The visual results of the 3D reconstructions are shown in Fig. 3. For all datasets, the resulting 3D models of the tunnel interior suffered from significant visual noise. Furthermore, the camera pose estimation failures led to spatially incorrect geometry in some areas. Navigating the viewpoint through these sections revealed a significant spatial warping effect. In contrast, when the same image datasets were processed in Metashape, image alignment failed for many images. Consequently, the reconstructed area was significantly smaller than that achieved with 3DGS, preventing the reconstruction of the tunnel's overall shape.

The results indicate that merely altering the video segmentation strategy is insufficient for achieving a high-precision 3D reconstruction of the tunnel space with 3DGS. However, 3DGS successfully estimated camera poses to some extent, whereas Metashape failed to align most of the images. This suggests that 3DGS exhibits greater robustness than conventional photogrammetry software for image-based 3D reconstruction in challenging environments such as tunnels. A key limitation in our current method is that preprocessing only involved distortion correction; the source footage contained moving vehicles, which were not removed. The presence of these dynamic objects is a likely contributor to the failures in camera pose estimation and the visual noise in the reconstruction. Therefore, future work will focus on improving reconstruction accuracy by implementing additional preprocessing steps, such as detecting and removing frames containing vehicles or inpainting the occluded areas.

4. Conclusions

This study investigated the feasibility of using 3DGS for the 3D reconstruction of tunnel spaces for maintenance and management applications. Our results demonstrate that achieving accurate 3D reconstruction of such environments cannot be accomplished merely by adjusting the image segmentation strategy from 360-degree video footage. Future work will focus on improving reconstruction accuracy by implementing preprocessing steps to address dynamic objects, for instance, by removing or inpainting vehicles from the source footage.

References

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, *42*(4).

Kagaya, Y., Otsuka, N., Katayama, M., Ishihama, S., Aoki, K., Ohtomo, Y., & Kawamura, Y. (2024). Accuracy Verification of Tunnel 3D Modeling Technology Using 3D Gaussian Splatting. *MMIJ 2024 Autum.*

Digital Life

Journal of Digital Life 2025, A-7

Conference Proceedings

Exploratory Study on Interpretation of Visual Information in Regional Disaster Management Plans Using Vision-Language Models

Ryuma Kawakubo [1], Kenji Nakamura [2], Kazuma Sakamoto [3] and Ryuichi Imai [4,*]

[1] Graduate School of Engineering and Design, Hosei University, 2-33 Ichigayatamachi, Shinjuku-ku, Tokyo162-0843, Japan
[2] Faculty of Information Technology and Social Sciences, Osaka University of Economics, 2-2-8 Osumi, Higashiyodogawa-ku, Osaka-shi, Osaka 533-8533, Japan
[3] Faculty of Production Systems Engineering and Sciences, Komatsu University, Nu 1-3 Shicho-machi, Komatsu-shi, Ishikawa 923-8511, Japan
[4] Faculty of Engineering and Design, Hosei University, 2-33 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843, Japan

1. Introduction

In Japan, local governments are mandated to formulate and, when necessary, revise Regional Disaster Management Plans annually to ensure effective disaster response. The revision process involves identifying the sections that require revisions from a vast amount of documentation, which demands considerable effort. In recent years, natural language processing techniques have been explored to support this process (Tomie et al., 2022). In particular, large language models capable of text generation and summarization can automatically identify and propose revisions based on the textual content of Regional Disaster Management Plans. However, these plans also contain content such as communication protocols and organizational structures during disaster scenarios. This information is presented in a variety of formats, including text, tables, and figures, and the formats differ across local governments. Processing such diverse content requires interpreting both textual and visual information. To address this, we focus on vision-language models (VLMs), which can interpret visual content. Previous studies have compared various methods of inputting tables and figures into VLMs (Zhou et al., 2025). However, whether current VLMs can effectively interpret complex visual formats remains unclear. These include flowcharts, detailed organizational charts, and tables containing dense information in a single cell. Moreover, the optimal input format for such figures and tables has not yet been clarified. Therefore, this study compares various VLMs and input formats to identify the most effective combination for interpreting tables and figures in Regional Disaster Management Plans.

2. Methods

In this study, we compare the answers of five VLMs to questions on tables and figures, using various input formats as shown in Fig. 1. The selected models support the Japanese language and demonstrate adequate performance on visual QA tasks. To control for performance differences due to model size, we limited our selection to models with around 7-8 billion parameters. As shown in Fig. 1(a), we compare three table input formats: HTML, Markdown, and image-based. The HTML and Markdown formats are generated using the document analysis model “yomitoku”. In Fig. 1(b), we compare two input formats for figures: an image with a title and 1-3 descriptive sentences, and one without any accompanying text.

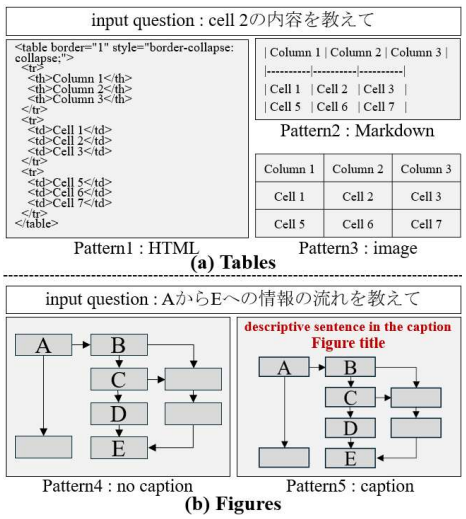


Fig. 1. Input formats and sample questions

3. Experiment and Results

This study focuses on tables in Regional Disaster Management Plans that describe deployment systems and criteria at the time of a disaster. The format in which this information is presented differs by prefecture. For example, in Miyazaki Prefecture, the table is presented as an image; in five prefectures including Fukui and Yamanashi, it is presented in plain

Published: 6 September 2025
* Correspondence: imai@hosei.ac.jp

text; and in the remaining 41 prefectures, it appears in tabular format. Accordingly, this study targets tables from 42 prefectures presented in non-plain-text formats. For tables, we use BERT Score as a quantitative evaluation metric to measure the similarity between the reference sentences and the generated answers. This metric reports Precision, Recall, and F-measure. Additionally, to supplement aspects not captured by numerical scores, such as semantic and contextual appropriateness, one author conducts a human evaluation using a four-point scale: ◎, ○, △, and ×. We extract one figure from the Regional Disaster Management Plan of each of Japan's 47 prefectures. Each figure is assessed by human evaluation on structural, terminological, and relational accuracy, using a binary scale of ○ and ×. Table 1 shows the BERT Score results for table-related questions. The highest F-measure, 0.674, was achieved by Qwen2.5-VL-7B-Instruct with Markdown input, while the lowest was from Llama-3-EvoVLM-JP-v2 with image input. Table 2 shows the results of human evaluation. Except for Qwen2.5-VL-7B-Instruct and sarashina2-vision-8b, all models failed to generate appropriate answers for table inputs. In particular, none of the models succeeded when using Markdown or image formats. For figures, Qwen2.5-VL-7B-Instruct and sarashina2-vision-8b generated correct answers in some cases but failed in most. In addition, the presence or absence of captions had no clear effect.

Table 1. Bert Score (Tables) ※Bold values in the table indicate the highest or lowest F-measure

models	Pattern1			Pattern2			Pattern3		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
llava-calm2-siglip [1]	0.549	0.560	0.554	0.543	0.661	0.594	0.549	0.560	0.554
Llama-3-EvoVLM-JP-v2 [2]	0.629	0.680	0.653	0.598	0.661	0.627	0.540	0.556	**0.547**
Llama-3-EZO-VLM-1 [3]	0.606	0.676	0.639	0.589	0.667	0.625	0.533	0.565	0.548
Qwen2.5-VL-7B-Instruct [4]	0.608	0.752	0.671	0.611	0.755	**0.674**	0.611	0.742	0.669
sarashina2-vision-8b [5]	0.566	0.690	0.621	0.568	0.695	0.624	0.597	0.673	0.630

1) https://huggingface.co/cyberagent/llava-calm2-siglip , 2) https://huggingface.co/SakanaAI/Llama-3-EvoVLM-JP-v2 , 3) https://huggingface.co/AXCXEPT/Llama-3-EZO-VLM-1
4) https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct , 5) https://huggingface.co/sbintuitions/sarashina2-vision-8b

Table 2. Human evaluation

models	Tables												Figures			
	Pattern1				Pattern2				Pattern3				Pattern4		Pattern5	
	◎	○	△	×	◎	○	△	×	◎	○	△	×	○	×	○	×
llava-calm2-siglip	2	2	0	38	0	2	0	40	0	0	0	42	0	47	0	47
Llama-3-EvoVLM-JP-v2	1	1	0	40	0	0	0	42	0	0	0	42	0	47	0	47
Llama-3-EZO-VLM-1	0	2	0	40	0	0	0	42	0	0	0	42	0	47	0	47
Qwen2.5-VL-7B-Instruct	8	24	1	9	10	20	0	12	9	22	1	10	13	34	15	32
sarashina2-vision-8b	1	7	1	33	2	7	0	33	3	4	3	32	7	40	6	41

Legend: Tables ◎:Complete and accurate ○:Accurate with extra info △:Partially incomplete ×:Incorrect or unrelated
Figures **I.** Structure: Matches flow **II.** Terminology: Consistent terms **III.** Relation: Correct arrows/connections
○:Satisfies criteria **I–III** ×:Fails to meet at least one of the criteria

4. Discussions

Llama-3-EvoVLM-JP-v2 and Llama-3-EZO-VLM-1 generated incorrect kanji characters and grammatical errors. This may be due to their base Llama model being primarily trained on English data, resulting in lower performance on Japanese-language tasks. In addition, all models except Qwen2.5-VL-7B-Instruct and sarashina2-vision-8b had difficulty interpreting images. A likely reason is that these models were trained mainly on images of natural scenes and objects, with limited training on charts and diagrams. This issue is particularly evident in figures and tables from Regional Disaster Management Plans. These often include dense text in table cells and complex flowcharts with arrows pointing in various directions, making them difficult to interpret.

5. Conclusions

This study confirms that Qwen2.5-VL-7B-Instruct is capable of interpreting tables and figures. On the other hand, answer generation largely depended on each model's ability to interpret visual information. These results suggest that fine-tuning may be necessary for accurate analysis of visual content in Regional Disaster Management Plans. Furthermore, even when visual content is difficult to interpret, its meaning can often be inferred from the surrounding textual context. Therefore, approaches that consider both visual elements and accompanying text should be explored.

References

Tomie, S., Hiroi, K. and Hatayama, M. (2022). Proposal and Evaluation of a Method for Supporting Issue Identification in Regional Disaster Management Plans Using Natural Language Processing Techniques. *IPSJ Technical Report*, IS-159 (4), 1-8

Zhou, W., Mesgar, M., Adel, H. and Friedrich, A. (2025). Texts or Images? A Fine-grained Analysis on the Effectiveness of Input Representations and Models for Table Question Answering. https://arxiv.org/pdf/2505.14131

Digital Life

Journal of Digital Life 2025, B-1

Conference Proceedings

Basic Research on Quantitative Analysis of Eight Stages of Shooting Using Skeletal Estimation Technology

Terumi Kakukawa [1], Kazuma Sakamoto [1]*, Yoshihiro Ueda [1], Iori Iwata [2] and Fuya Shibata [2]

[1] Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[2] Graduate School of Sustainable Systems Science, Komatsu University, Komatsu, Ishikawa 923-8511, Japan

1. Introduction

According to the medium-term plan (All Japan Kyudo Federation al., 2024) for its 80th anniversary, it can be said that the majority of registrants are beginners and intermediates. As methods of checking shooting form, reviewing one's posture in a mirror or recording videos is often used. However, for beginners, it is difficult to independently identify the specific points requiring correction. Ideally, advanced practitioners or experts should provide feedback and advice. Nevertheless, there is a shortage of young high-ranking practitioners, and in school club activities, teachers are often required to act as advisors even without prior experience. In recent years, researches have been conducted to visualize Kyudo shooting form using image recognition techniques (Morozumi et at., 2018) (Kawase et al.,2011). However, these researches present several limitations: the analysis is restricted to the upper body, quantitative evaluation of shooting techniques has not been performed, the subjects are predominantly advanced practitioners with no evaluation of beginners, and the lack of normalization introduces variations among individuals. To address these issues, the present research applies pose estimation to reference images and videos of both beginners and advanced practitioners captured with a smartphone. Through quantitative analysis and discussion of these data, this research aims to contribute to the improvement of beginners' skills.

2. Methods

The method estimates pose and joint angles with MMPose from front-view videos of Kyudo. Movements from Dozukuri to Hanare in the Shahō Hassetsu are analyzed, excluding the head. Videos are split into frames; for each stage, the terminal frame (just before transition) is used. From detected keypoints, we compute the centers of the waist, shoulders, and ankles, shift the coordinate origin to the waist center, and normalize all coordinates by the waist–shoulder distance. Using Excel's atan2, we calculate waist and shoulder horizontality, leg verticality, and torso verticality, converting to degrees. Target values reflecting Sanjujūmonji are ≈180° (waist), ≈180° (shoulders), ≈90° (waist-to-ankle line), and ≈270° (torso). Left-side keypoints define the reference; +y is downward. For multiple shots, per-subject averages are compared to a reference form.

The comparison between the reference model and each participant was conducted by calculating the angular differences for four key postural parameters. Experimentally, four shots each by a Kyoshi 6-dan and two unranked beginners were analyzed and averaged. As the reference model, All Japan Kyudo Federation Shahō Hassetsu images were used, with backgrounds blacked out to improve pose-estimation robustness. Angles for the reference and participants were computed via the proposed pipeline and compared.

3. Results and Discussion

As demonstrated in Fig 1, a sample frame following pose estimation for Shahō Hassetsu (Dozukuri) is presented.
As demonstrated in Table 1, the initial model (an advanced practitioner's Dozukuri) demonstrates the estimation of angular values. In contrast, the results presented in Table 2 demonstrate the angular estimation of two Dozukuri models employed by novices. As illustrated in Fig 1, Beginner 1 exhibited a greater degree of shoulder tilt in comparison to the other participants, a phenomenon that is distinctly evident in the angle values enumerated in Table 2. The present research demonstrated that quantitative comparisons between the reference model and participants enable the identification of technical improvement points through angle difference analysis. The advanced practitioner

Published: 6 September 2025
* Correspondence: kazuma.sakamoto@komatsu-u.ac.jp

produced values close to the reference across additional phases. For the beginners, correction points became explicit in the metrics, for example large shoulder angles and increased torso inclination in certain phases. However, some phases for the reference model did not produce the expected values. Participant angles were obtained by averaging four shots at the terminal frame of each phase, whereas the reference was computed from a single image. As a result, the reference values likely include offsets from pose estimation.

Reference Model Advanced practitioner Beginner 1 Beginner 2

Fig 1. Results after skeletal estimation

Table1. Reference Model and Advanced practitioner's Dozukuri angle estimation results

Reference Model	Angle [Degrees]			
Times/Case	Waist	Shoulders	Waist to Ankle	Torso
1	180.0	180.0	90.7	271.0
Advanced Practitioner	Angle [Degrees]			
Times/Case	Waist	Shoulders	Waist to Ankle	Torso
1	184.8	177.1	90.0	270.0
2	184.8	180.0	90.0	270.0
3	180.0	177.1	91.4	270.0
4	180.0	180.0	90.7	271.9
Average	182.4	178.6	90.5	270.5

Table2. Beginner's Dozukuri angle estimation results

Beginner 1	Angle [Degrees]			
Times/Case	Waist	Shoulders	Waist to Ankle	Torso
1	180.0	189.0	91.9	271.1
2	180.0	186.0	91.3	268.9
3	180.0	188.5	91.3	269.0
4	180.0	188.5	93.2	268.9
Average	180.0	188.0	91.9	269.5
Beginner 2	Angle [Degrees]			
Times/Case	Waist	Shoulders	Waist to Ankle	Torso
1	180.0	180.0	90.6	271.0
2	180.0	180.0	90.6	270.0
3	180.0	177.0	90.0	270.0
4	180.0	177.0	90.0	270.0
Average	180.0	178.5	90.3	270.3

4. Conclusions

This research conducted a quantitative analysis of Kyudo shooting form using pose estimation and provided interpretive insights. Although the estimated angles have not yet been validated against ground truth, the approach quantified differences in movement among participants and identified concrete correction points for beginners. As future work, we will first verify that the estimated angles match the actual physical angles. We will also expand the set of reference data to enable more appropriate comparisons. Additionally, establishing statistically validated thresholds for distinguishing between measurement error and meaningful postural differences will be essential for more reliable technical assessment. Furthermore, by integrating the present results with the distribution of arrow landing positions, we expect to enable quantitative analyses that more directly support improved shooting accuracy.

References

All Japan Kyudo Federation. (2024). the medium-term plan of All Japan Kyudo Federation 2023-2029. https://www.kyudo.jp/aboutus/plan.html

Morozumi, T., Ozono, T., & Shintani, T. (2018). Training data on Japanese archery motion recognition for self-training support. *In Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence* (JSAI2018) 1-2

Kawase, H., & Matsuzuka, A. (2011). A support system for learning sports by using sound and image processing in real time. In Interaction 2011 — *Proceedings (Information Processing Society of Japan Symposium Series*, (3), 465-468

Digital Life

Journal of Digital Life 2025, B-2

Conference Proceedings

Research on verifying effectiveness of SEO measures estimated from search results

Sho Okado[1], Masaya Nakahara[1] and Sakamoto Kazuma[2]

[1] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Production Systems Engineering and Sciences, Komatsu University, 1-3 Yonchomachi-nu, Komatsu-shi, Ishikawa 923-8511 Japan

1. Introduction

With the advancement of web production technology, a vast number of web pages continue to be produced daily worldwide. Regardless of whether they are individuals or businesses, websites generate advertising revenue from users who visit them and sell goods on e-commerce sites. Since users typically view websites that rank highly in web search results, it is crucial for site operators—for whom the number of visitors directly impacts income—to ensure their websites appear at the top of search rankings. Therefore, website operators must understand the evolving evaluation criteria of web search engine ranking algorithms and implement SEO (Search Engine Optimization) measures on their own websites accordingly. The authors previously proposed a method to derive effective SEO factors based on historically collected data of ranking fluctuations and web page content changes from existing research (Tanaka, S. et al, 2024). They confirmed that applying these effective SEO measures to historical web pages improved their rankings. However, they have not yet confirmed whether applying these measures to actual websites improves rankings. Therefore, this study aims to verify the usefulness of the existing method by confirming whether applying the highly effective SEO measures estimated using the existing method leads to improved rankings.

2. Methods

This study uses existing method (Tanaka, S. et al, 2024) to estimate SEO-effective items within the same genre as the target web pages. It then verifies whether applying previously unimplemented items improves rankings. The newly applied countermeasures are shown in Table 1. These items are limited to those within the top third of countermeasures ranked by existing methods that have not yet been applied.

Fig.1. Applied SEO measures.

Set the web page display speed to 6 seconds or less	Set the DOM size to 1500 or less	Set the og_site_name meta tag
Set the page display speed to 6 seconds or less and configure a button with a recognizable name.	Increase the number of times the first query is used within text other than A tags	Set the character count for Description to 120 characters
Increase the character count of text excluding anchor text	Set the viewport	Set the second query in the H1 tag
Increase the number of characters in the text	Include a Doctype declaration in the HTML file	Set the Title character count to 40 characters or fewer
Adaptive Text Compression	Set the second query in the Title	Increase the number of times the second query is used within the text
Set different text for the Title and H1 tags	Set the og_image meta tag	Set the second query in the Description
Set the initial display time for text and images to 4 seconds or less.	Set the og_type meta tag	Set the first query as the Title
Increase the number of times the second query is used within text other than A tags	Set the og_url meta tag	Increase the number of alt tags containing the second query
Set alt tags for linked images	Set the og_description meta tag	—

Published: 6 September 2025
* Correspondence: nakahara@oecu.jp

3. Demonstration experiment

This experiment examines the ranking fluctuations before and after applying SEO measures to Web Pages A and B, which explain SEO strategies. Google was selected as the search engine for evaluation. The evaluation period began monitoring ranking changes from early May 2025, when the SEO measures were applied, through the end of July 2025. Ranking changes were checked every 10 days over this three-month period. However, monitoring of ranking changes failed around July 10, 2025, so measurements were not possible for that period. This interval is represented by a dotted line using linear interpolation. The ranking change results are shown in Fig. 1. Fig. 1 shows that Page A entered the top 100 search results on June 12, 2025, and Page B did so on June 26, 2025. Page A reached a peak position of 38th, while Page B peaked at 62nd. This is likely because it took time for the crawler to traverse web pages in similar genres to these pages and complete the assignment of correct indexes. From the above, it was confirmed that applying SEO countermeasures obtained through existing methods leads to improved web page rankings, thus verifying the usefulness of existing methods.

However, the rankings have not improved to the top 10 positions on the first page of search results. Therefore, further investigation is needed to determine whether applying other optimization items could improve rankings and to identify any factors beyond the SEO optimization items. Additionally, when applying these SEO optimization items to the web pages this time, the update content was conceived and applied by the web page administrators, requiring a significant amount of time to create the update pages. Furthermore, since it is impossible to predict which SEO countermeasure items will lead to ranking improvements and to what extent, the only approach is to apply high-priority countermeasures and verify their effectiveness. Therefore, it is considered necessary to develop a system that supports the efficiency of web page administrators' update tasks. This system would allow administrators to specify the application of SEO countermeasure items via checkboxes, etc. Upon specification, it would automatically use models like GPT or other large language models (LLMs) to simultaneously display proposed changes for the updated web page, along with the predicted ranking and the estimated contribution to ranking improvement.

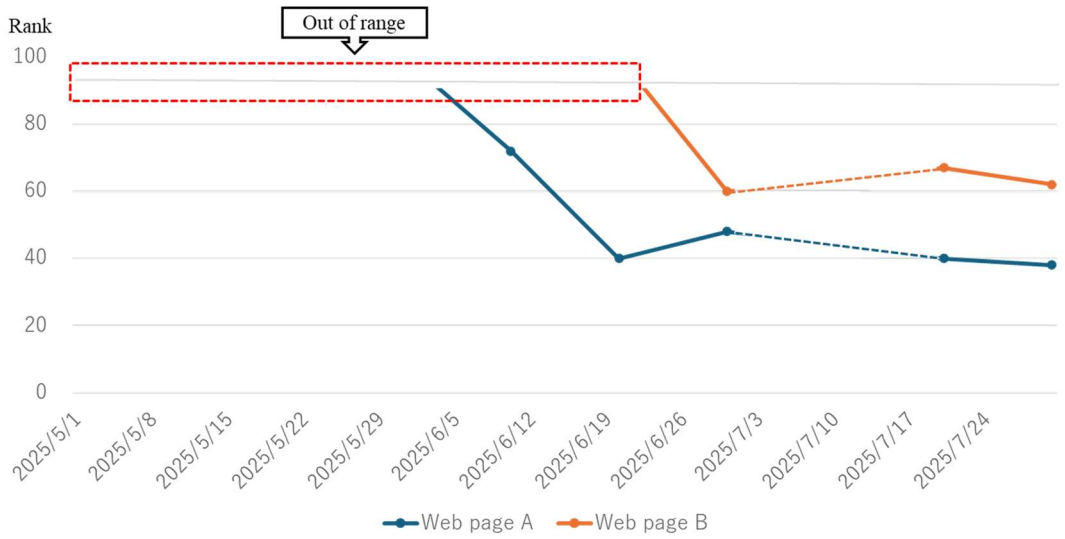


Fig.1 Results of logarithmic complexity and coherence for each genre.

4. Conclusions

This study verified the effectiveness of existing methods by applying top-ranked SEO items—derived from a group of web pages in the same genre—to web pages explaining SEO strategies, to determine if this would improve their rankings. The results of the empirical experiment showed that web pages ranked below 100th place improved to as high as the 30s within three months, confirming the usefulness of existing methods. Moving forward, to achieve further ranking improvements, we plan to explore applying other SEO items and investigating other ranking factors. Additionally, we intend to develop a system to streamline the process of implementing SEO items onto web pages.

References

Tanaka, S., Nakamura, K., Teraguchi, T., Yamamoto, Y., Sakamoto, K., Nakahara, M., Kususmoto, N., Iwamoto, T., (2024). Research for Quality Evaluation of Web Pages Focused on SEO Internal Measures, *Journal of Japan Society of Civil Engineers for Artificial Intelligence and Data Science*, *5*(1), 269-280.

Digital Life

Journal of Digital Life 2025, B-3

Conference Proceedings

Research on relationship between categories of posted videos using LDA

Kyoya Takiguchi[1], Masaya Nakahara[2] and Kazuma Sakamoto[3]

[1] Graduate School of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[2] Faculty of Information Science and Arts, Osaka Electro-Communication University, 1130-70 Kiyotaki, Shijonawate-shi, Osaka 575-0063, Japan
[3] Faculty of Production Systems Engineering and Sciences, Komatsu University, 1-3 Yonchomachi-nu, Komatsu-shi, Ishikawa 923-8511 Japan

1. Introduction

In recent years, video sharing platforms such as YouTube have expanded rapidly, with the number of videos posted and viewing time continuing to increase. In such an environment, the importance of recommending technologies that efficiently present content tailored to user preferences and technologies that predict the future evaluation of videos is increasing. Videos have categories that are officially defined by the platform, and the classification criteria are completely clear. However, despite this formal classification, considering actual user viewing behavior and content characteristics, there may be potential relationships between categories. Clarifying such relationships is expected to provide fundamental insights that contribute to improving the accuracy of recommendation models and evaluation predictions. Existing research (Sakamoto, K., et al. 2020) has also widely analyzed the use of LDA (Latent Dirichlet Allocation) (David, B., et al. 2003) has been widely used for topic extraction in microblogs and news articles, as well as for analyzing the genre structure of television programs, and has been applied to a variety of media data. These studies have shown that capturing the underlying latent topic structure behind a clearly defined classification system is effective for information recommendation and content understanding. However, systematic analyses of the relationships between categories on video-sharing platforms remain limited. Therefore, this study aims to investigate the latent relationships between categories of uploaded videos using LDA, establish criteria for training data used in evaluation prediction, and gain insights that contribute to improving the accuracy of future video evaluation prediction models.

2. Methods

In this chapter, we will explain the procedure for investigating the relationships between categories of posted videos. First, we transcribe the uploaded videos into text data. The tool used for transcription is ReazonSpeech, an open-source speech recognition model specialized for Japanese. Next, we use LDA to extract topics from the converted text data. LDA is a method that statistically estimates the relationship between documents and topics based on the frequency of word occurrences in a text. By converting posted videos into text and applying LDA, it becomes possible to numerically analyze the composition of topics contained in each video. This allows us to quantitatively understand the topic composition of each category by aggregating and comparing topic distributions at the category level. Furthermore, by analyzing the existence of common topics and similarities in topic ratios between different categories, we can extract semantic relationships that are difficult to capture through superficial labeling.

3. Demonstration experiment

In this experiment, we applied LDA to actual posted videos and verified the potential relationships between categories. The genres of the videos we targeted were game commentary videos, and the game genres were “FPS,” “RPG,” “sandbox,” and “horror.” We targeted 15 videos for each game genre. We estimated the number of topics in the LDA model by varying it from 1 to 200. This allowed us to comprehensively examine the trends of topics extracted for each category and the potential relationships between genres. To confirm the validity of the model, we used coherence scores to measure topic consistency and log-perplexity to indicate model fit as metrics. By referencing these metrics,

Published: 6 September 2025
* Correspondence: nakahara@oecu.jp

we determined the appropriate range of topics to use in the analysis and identified clues to clarify the potential relationships between categories. As shown in Fig. 1, it was confirmed that increasing the number of topics consistently decreased the logarithmic complexity and improved the model fit in all genres. On the other hand, coherence shows differences in the degree of increase and stability across genres, which is believed to reflect the characteristics of each genre. When examining each genre individually, Sandbox has a very wide range of player actions and playstyles, resulting in high freedom and topic dispersion, with coherence remaining around 0.24 and showing significant fluctuations. The optimal number of topics is estimated to be around 80–120. RPGs often unfold along the framework of a story or progression, making topics naturally cohesive and showing relatively stable increases, reaching around 0.28, with 100 being an appropriate range. FPS had the highest coherence, exceeding 0.32. Due to the limited patterns of gameplay and terminology, topics are clearly divided, enabling stable topic formation. The optimal number of topics is most easily achieved in the range of 100 to 150. Horror games have a wide range of expression forms, with different works focusing on psychological horror, exploration, action elements, etc. Therefore, topics tend to be dispersed, with coherence at around 0.22 and significant fluctuations. The optimal number of topics is around 50 to 100, and increasing the number of topics beyond this point only improves interpretability to a limited extent. Overall, it is considered most appropriate to set the number of topics to around 100 for all genres. In FPS and RPG, around 100 topics show stable and high coherence, making it easier to ensure consistency in topics. On the other hand, in sandbox and horror, there is a large dispersion of topics, and increasing the number of topics only leads to limited improvement in consistency. However, using around 100 as a benchmark still allows for a certain degree of interpretability. When comparing genres, FPS and RPG tend to have topics that coalesce based on “rules or frameworks,” while sandbox and horror genres exhibit contrasting characteristics where topics tend to vary due to “diversity.” Therefore, while the nature of each genre differs, a topic count of around 100 remains the most appropriate benchmark across all genres.

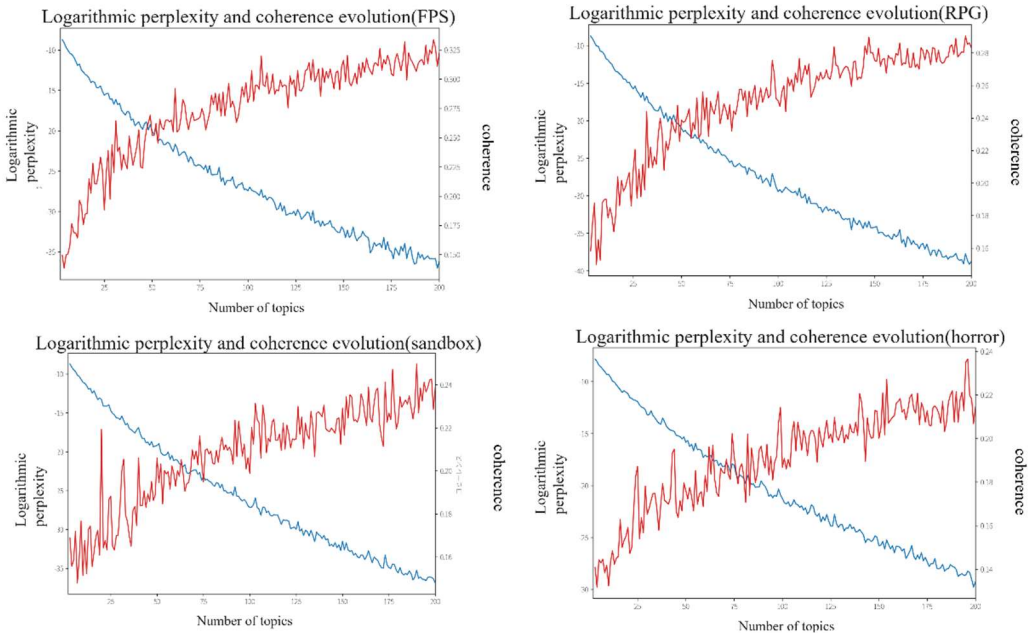


Fig.1 Results of logarithmic complexity and coherence for each genre.

4. Conclusions

In this study, we used LDA to investigate the relationships between categories and the optimal number of topics. Through verification experiments, we found that topics tend to be more cohesive in FPS and RPG games, while they tend to be more scattered in sandbox and horror games. We also confirmed that the optimal number of topics is around 100 for all genres. Going forward, we plan to build a video evaluation prediction model based on the results of this study.

References

David, B., Andrew, N., Michael, J., (2003). Latent Dirichlet Allocation, *The J. of Machine Learning Research,* 993-1022.

Sakamoto, K., Nakamura, K., Yamamoto, Y., Tanaka, S & Nakamura T. (2020). Practical Study Regarding Social Sensing Technologies for Extracting Unordinary Phenomena Considering User Attributes with Focus on Different Behavior from Normal Time, *journal of Japan Society for Fuzzy Theory and Intelligent Informatics.*, 556-569.

Digital Life

Journal of Digital Life 2025, B-4

Conference Proceedings

Text Mining on Benefits and Challenges of AI-Generated Academic Paper Short Videos and Text Summaries

Hayato Sezaki [1,*], Takashi Goto [2], Ayako Kurono (Fukunaga) [3], Hideo Kawamata [4] and Kayoko Kurita [1,5]

[1] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8563, Japan
[2] IBM Japan, 2-6-1 Toranomon, Minato, Tokyo 105-5531, Japan
[3] The United Graduate School of Agricultural Science, Iwate University, 3-18-8 Ueda, Morioka, Iwate 020-8550, Japan
[4] Graduate School of International Cooperation Studies, Kobe University, 1-1 Rokkodai, Nada, Kobe, Hyogo 657-8501, Japan
[5] Center for Research and Development of Higher Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

1. Introduction

Accessing and understanding academic literature remains challenging for non-experts. While generative AI services are being developed to support this process, few utilize image or video generation. Sezaki et al. (2024) proposed "Paper 2 Clip," a generative AI system that converts academic papers into short videos, inspired by the popularity of vertical video among youth. Previous analysis (Sezaki et al., in press) found that users consistently rated short videos more favorably than text summaries, regardless of user attributes. To clarify user perspectives, this study uses text mining to examine perceived benefits and challenges of AI-generated academic paper short videos and text summaries.

2. Methods

A web survey via QiQUMO was conducted on Feb 27–28, 2025, with 517 valid responses from Japanese users aged 15–69. Participants viewed two short videos and read two text summaries (each 1 minute), generated from four English papers published in the *Journal of Digital Life* (CC BY 4.0), using ChatGPT, Sora, and other tools. After each viewing, participants rated the content using nine 6-point Likert items (e.g., Relevance, Interest, Understanding). Open-ended responses were analyzed using co-occurrence network analysis in KH Coder, excluding entries such as "none." Based on mean impression scores, users were categorized into Low, Middle, or High groups as external variables (Table 1).

Table 1. Distribution of mean Likert scale scores used for external variables

Format	Low Group (≤ 33.3%)	Middle Group (33.3% – 66.6%)	High Group (≥ 66.6%)	Overall Mean ± SE (N = 517)
Short Videos	1.28 – 2.94	3.00 – 3.56	3.61 – 5.78	3.27 ± 0.036
Text Summaries	1.11 – 2.78	2.83 – 3.44	3.50 – 5.56	3.16 ± 0.036

Note. The 9-item scale showed high reliability (Cronbach's α = .881 for short videos; .883 for text summaries).

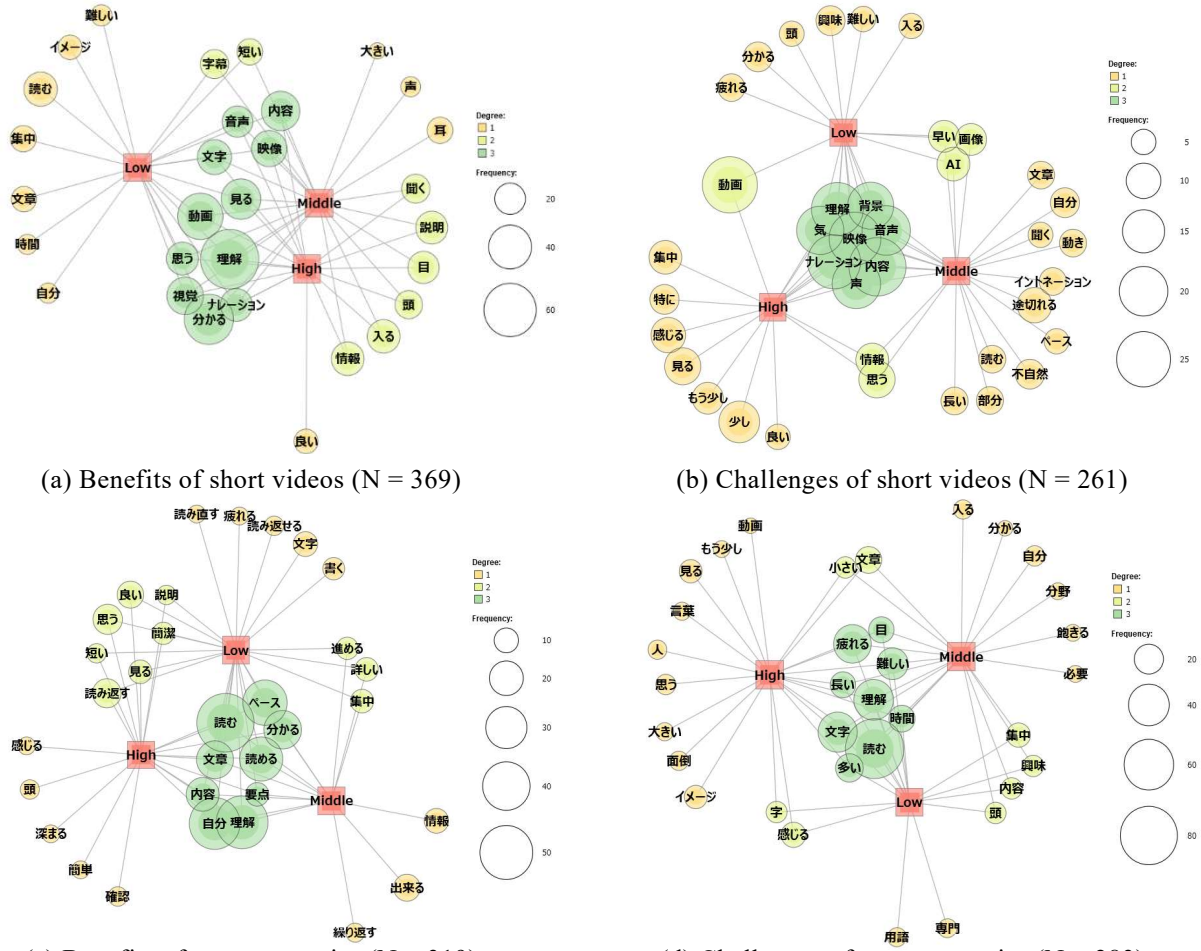
3. Results

Fig. 1 shows co-occurrence networks. Node size indicates word frequency. Node color reflects the degree of group overlap: degree 1 (orange) = appeared only in one group, degree 2 (yellow) = appeared in two groups, and degree 3 (green) = appeared in all three groups. Red squares indicate group labels from Table 1.

In the benefits of short videos (Fig. 1a), green-colored nodes indicate words common across all three groups, such as "video (動画)," "visual (視覚)," "image (映像)," "text (文字)," "audio (音声)," "narration (ナレーション)," "content (内容)," "understanding (理解)," "understand (分かる)," and "think (思う)," suggesting that the multisensory, multimedia format unique to short videos may have contributed to content comprehension. In the challenges of short videos (Fig. 1b), similarly shared terms included "narration (ナレーション)," "audio (音声)," "voice (声)," "background (背景)," "image (映像)," "content (内容)," "understanding (理解)," and "attention (気)," indicating discomfort with AI-generated narration and unrelated background visuals that disrupted concentration.

For text summaries, common benefits (Fig. 1c) included "read (読む)," "pace (ペース)," "self (自分)," "understanding (理解)," "text (文章)," and "key points (要点)," showing that participants valued the ability to read structured content at their own pace. Common challenges (Fig. 1d) included "read (読む)," "text (文字)," "many (多い)," "long (長い)," "difficult (難しい)," and "tiring (疲れる)," with "technical term (専門用語)" also noted in the Low group, indicating that reading burden, difficulty, and too many technical terms were challenges.

Published: 6 September 2025
* Correspondence: hsezaki@s.h.k.u-tokyo.ac.jp



(a) Benefits of short videos (N = 369)

(b) Challenges of short videos (N = 261)

(c) Benefits of text summaries (N = 319)

(d) Challenges of text summaries (N = 283)

Fig. 1. Co-occurrence network of benefits and challenges of short videos and text summaries

4. Discussions

Short videos offered multimodal benefits, including visuals, narration, and subtitles, which enhanced user engagement. However, some users expressed discomfort with AI-specific elements such as synthetic voices and unrelated visuals. Some videos featured nonsensical graphs or inaccurate system diagrams that disrupted comprehension. Text summaries were valued for allowing self-paced reading, but even brief texts often included technical terms from the original papers, which some users found difficult. These findings highlight the importance of reducing hallucinations, using multimedia features effectively, and limiting technical terms in AI-generated academic paper short videos.

5. Conclusions

This study analyzed user impressions of AI-generated academic paper short videos and text summaries, based on English-language journal articles, using text mining of open-ended responses. While the findings are based on subjective impressions, they offer valuable insights into the benefits and challenges of each format. As many participants had limited experience with reading academic papers, future research should target users more familiar with scholarly literature. Additionally, further work is needed to explore practical implementation in educational and research settings, including university classes, academic conferences, and workshops.

Acknowledgements

This study was supported by GSFS, UTokyo through a Challenging New Area Doctoral Research Grant (C2410), and by JST BOOST (JPMJBS2418). We thank the *Journal of Digital Life* for providing access to the source articles.

References:

Sezaki, H., Goto, T., Kurono, A., and Kawamata, H. (2024). “Paper 2 Clip” – A Generative AI System for Converting Academic Papers to Short Video Clips –, *The Conference of Digital Life vol. 2*, Digital INSPIRE, DI-1.

Sezaki, H., Goto, T., Kurono, A., Kawamata, H., Kurita, K. (In press). Comparing User Perceptions of AI-Generated Short Video Clips and Text Summaries of Academic Papers: Toward the Development of the “Paper 2 Clip” System, *Procedia Computer Science,* KES 2025, volume and page numbers TBD.

Digital Life

Journal of Digital Life 2025, B-5

Conference Proceedings

Fundamental Study on Indoor Space Representation Using Local Spatial IDs Derived from Mobile Terminal Point Clouds

Ryo Komiya [1], Kenji Nakamura [2], Yoshinori Tsukada [3], Yoshimasa Umehara [4], Yasuhito Niina [5] and Ryuichi Imai [6]

[1]Doctoral Course Graduate School of Engineering and Design, Hosei University, 2-33 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843, Japan
[2]Faculty of Information Technology and Social Sciences, Osaka University of Economics, 2-2-8 Osumi, Higashiyodogawa-ku, Osaka-shi, Osaka 533-8533, Japan
[3]Faculty of Engineering, Reitaku University, 2-1-1 Hikarigaoka, Kashiwa-shi, Chiba 277-8686, Japan
[4]Faculty of Business Administration, Setsunan University, 17-8 Ikedanakachō, Neyagawa-shi, Osaka 572-8508, Japan
[5]Asia Air Survey Co., Ltd., 1-2-2 Manpukuji, Asao-ku, Kawasaki-shi, Kanagawa 215-0004, Japan
[6]Faculty of Engineering and Design, Hosei University, 2-33 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843, Japan

1. Introduction

In Japan, urban-scale digital twins are being developed to enhance national resilience by addressing challenges in urban planning, disaster prevention, and infrastructure management (Cabinet Secretariat, Government of Japan, 2023). Publicly available 3D point cloud data, hereafter referred to as "point cloud"—such as those from the PLATEAU project led by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT, 2020)—are central to these efforts. Recently, the Spatial ID system — which assigns unique voxel-based identifiers to locations on Earth — has been proposed, further accelerating the development of digital twins. However, the indoor point cloud remains scarce due to the cost and privacy concerns of high-end surveying equipment. In contrast, mobile terminals like smartphones now allow affordable acquisition of such data, though typically captured in independent local coordinate systems, which poses challenges for integration with georeferenced outdoor data. The Local Spatial ID system has been proposed as a solution. It facilitates spatial referencing in localized environments, such as building interiors, with customizable voxel sizes and zoom levels. Moreover, it facilitates integration across distinct spatial environments, enabling interoperability with globally referenced datasets (Ministry of Economy, Trade and Industry et al., 2025). This study proposes a method for linking indoor point cloud—managed in local coordinate systems—to urban-scale digital twins using the Local Spatial ID specification. This approach may enhance the efficient integration of point clouds from mobile terminals into broader urban-scale digital twins.

2. Methods

In the proposed method, indoor environments are subdivided into manageable local spaces, such as rooms or floor units, each represented as a cuboid with a defined origin, dimensions, and azimuth. Based on the Real Estate ID Rule Guidelines (MLIT et al., 2022), each space is assigned a unique four-digit code based on the floor or room number. For example, as shown in Table 1, G002 indicates a second-floor corridor, and G201 refers to Room 201 connected to G002. Point clouds captured by mobile terminals are aligned within each space's local coordinate system, then converted into Local Spatial IDs, which consist of a zoom level (z) and index values (f, x, y) along the z/f/x/y axes. In this study, zoom level 9 is used, corresponding to a voxel size of 0.10 m. Integration across spaces is managed using connection metadata, which includes space codes and Spatial IDs of

Table 1. Example of Node Connection Information

Target code	Source code	Target id (z/f/x/y)	Source id (z/f/x/y)
Global	G002	27/87/119206578/52845384	9/2/234/129
G002	G201	9/2/587/314	9/2/9/19

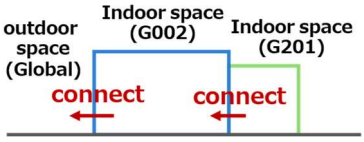


Fig. 1. Overview of Spatial Connections

Published: 6 September 2025
* Correspondence: imai@hosei.ac.jp

connection points (Table 1). An overview of node connections is shown in Fig. 1. Connections labeled "Global" represent links between local spaces and the global coordinate system, while entries such as "G002" indicate local-to-local connections. Importantly, connections can span Spatial IDs with differing zoom levels. Using this metadata, the global coordinates of junction points can be calculated. Consequently, the origin of a local space can be mapped onto the global coordinate system using its azimuth and Spatial ID. Furthermore, chaining local-to-local connections enables indirect integration with globally referenced data.

3. Results

This chapter visualizes Local Spatial IDs generated from indoor point clouds to evaluate the proposed method. A total of 33 point clouds, captured using an iPhone 13 Pro Max, were used for validation. Fig. 2 shows a portion of the generated voxel model with a voxel size of 0.10 m. Its origin, set at (-12.35, -0.73, -0.15), lies within a local coordinate system. Connection metadata was defined at an entrance linking a second-floor corridor to the outdoor space (Table 1). The Global Spatial ID was set at zoom level 27, corresponding to a voxel size of approximately 0.12 m. The coordinates used in the connection metadata were defined in a plane rectangular coordinate system as (-8721.85, -33764.58, 21.99). Fig. 3 presents the visualization result of the Local Spatial IDs on the global coordinate system using the connection metadata. The voxel model based on Local Spatial IDs is overlaid with the PLATEAU building model and airborne LiDAR point clouds, both of which are globally referenced. As shown in the figure, the voxel model is correctly positioned, demonstrating the potential of Spatial IDs for integrated management of local and global coordinate systems.

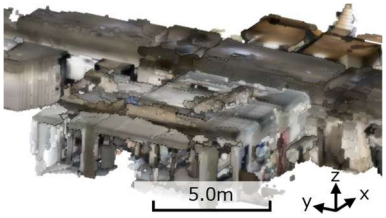


Fig. 2. Generated Voxel Model (Voxel Size: 0.01 m)

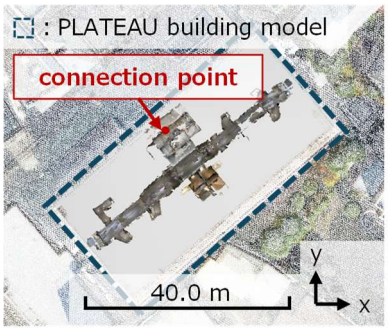


Fig. 3. Visualization Result of Local Spatial IDs on the Global Coordinate System

4. Discussions

As demonstrated in Chapter 3, Local Spatial IDs generated from point cloud captured by mobile terminals were successfully visualized within a global coordinate system. This confirms that indoor spaces can be subdivided, processed independently, and later integrated using Spatial ID-based methods. However, since the connection metadata used Spatial IDs at zoom level 27, this may result in positional errors of up to 0.12 m. To improve accuracy, it will be necessary to define multiple connection points or adopt alternative error-tolerant strategies. Also, as shown in Fig. 4, geometric discrepancies between the PLATEAU model and the generated indoor data suggest that the indoor space may not be directly connected to the outdoor environment. To address this issue, the geometry of outdoor areas near connection points should be modeled in greater detail, enabling the creation of more realistic digital twins that seamlessly integrate indoor and outdoor spaces.

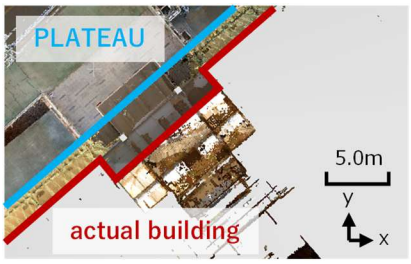


Fig. 4. Enlarged View of the Entrance Area

5. Conclusions

This study proposed a method for integrating 3D geometry obtained from point clouds in local coordinate systems into a global coordinate system as a means to construct digital twins of indoor spaces. The effectiveness of the method was demonstrated by successfully visualizing Local Spatial IDs—recorded in local coordinates—within the global coordinate system, using custom-defined connection metadata. Future work will focus on leveraging the portability and ease of point cloud acquisition offered by mobile terminals to generate detailed models of building exteriors, with the ultimate goal of constructing digital twins that seamlessly integrate indoor and outdoor environments.

References

Cabinet Secretariat, Government of Japan. (2023). *Fundamental Plan for National Resilience against Natural Disasters*, https://www.cas.go.jp/jp/seisaku/kokudo_kyoujinka/pdf/kk-honbun-r057028.pdf

Ministry of Economy, Trade and Industry et al. (2025). *Spatial ID Guidelines for the Utilization of Four-Dimensional Spatiotemporal Information (Version 1.0)*, https://www.ipa.go.jp/digital/architecture/Individual-link/nl10bi000000377d-att/4dspatio-temporal-id-guideline.pdf

Ministry of Land, Infrastructure, Transport and Tourism. (2020). *PLATEAU*, https://www.mlit.go.jp/plateau/

Ministry of Land, Infrastructure, Transport and Tourism et al. (2022). *Real Estate ID Rules Guidelines*, https://www.mlit.go.jp/tochi_fudousan_kensetsugyo/content/001769626.pdf

Digital Life

Journal of Digital Life 2025, B-6

Conference Proceedings

State changes when tick size is changed

Hiroyuki Maruyama [1]

[1] Faculty of Commerce Takushoku University, 4-14 Kohinata 3-chome, Bunkyoku, Tokyo 112-8585, Japan

1. Introduction

Institutional design is important in securities markets. One such initiative is setting tick size. Tick size represents the unit by which an investor can change the order price when buying or selling assets in the securities market. Changing this can provide investors with a variety of conveniences, such as increasing their options.

The purpose of this study is to investigate the timing of the impact of tick size changes. In particular, we will analyze the tick size change implemented by the Tokyo Stock Exchange, Japan's leading stock exchange. This was carried out in 2014, and two changes were made (referred to as Phase 1 and Phase 2, respectively).

2. Methods

The sample stocks used in this study were selected based on the following criteria: Among the TOPIX 100 stocks, Phase 1 focused on stocks priced above ¥3,000, and Phase 2 focused on stocks priced below ¥5,000. In addition, stocks must always be trading, their stock prices must be within the target price range for the tick size change, and stocks priced below ¥100 or above ¥30,000 were excluded (Ahn et al., 2007).

Data were collected for three months before and after the change (Ahn, Cao, and Choe, 1996).

All data were obtained from the JPX Data Cloud.

The model used in the analysis is a threshold autoregressive model. This was developed by Tong and Lim (1980) and Tong (1983). This model takes different parameters depending on the relationship between the threshold and the state variable. Specifically, it is expressed by the following equation.

$$y_t = \begin{cases} \alpha_0^0 + \beta_1^0 y_{1-1} + \cdots + e_t \\ \alpha_0^1 + \beta_1^1 y_{1-1} + \cdots + e_t \end{cases} \quad (1)$$

Here, α_0^i and β_t^i represent coefficients, and e_t represents the error.

In this study, estimation was performed using y_t as the logarithmic average trading volume. In doing so, the stocks subject to the tick size change were classified by industry, and estimation was performed for each industry.

3. Results

The sample size is shown in Table 1.

A summary of the estimated thresholds is presented in Table 2.

Published: 6 September 2025
h-maruya@ner.takushoku-u.ac.jp

Table 1. Summary of sample size

	Fisheries, Agriculture and Forestry	Mining Industry	Construction industry	Manufacturing industry	Electricity and Gas Industry	Transportation and Information and Communications	Commercial	Finance and Insurance	Real Estate	Service Industry
phase 1		2	1	20		6	1	2	2	2
phase 2	1	3	4	39	4	4	7	14	3	

Table 2. Summary of estimation results

	Fisheries, Agriculture and Forestry	Mining Industry	Construction industry	Manufacturing industry	Electricity and Gas Industry	Transportation and Information and Communications	Commercial	Finance and Insurance	Real Estate	Service Industry
phase 1		0.81	0.16	0.17		0.22	0.33	0.15	0.65	0.28
phase 2	0.70	0.75	0.39	0.78	0.28	0.84	0.84	0.78	0.78	

4. Discussions

In this study, we attempted to detect the timing of the impact of the tick size change.

The results are summarized in Table 2. As can be seen, in Phase 1, industries were able to be divided into two groups: those whose status changed before the change (threshold less than 0.5) and those whose status changed after the change (threshold higher than 0.5). Similarly, in Phase 2, industries were able to be divided into two groups: those whose status changed before the change and those whose status changed after the change.

5. Conclusions

In this study, we analyzed the timing of the impact of tick size changes. As a result, we found that there were two types of industries: those whose conditions changed before the change and those whose conditions changed after the change.

In future research, we would like to conduct more detailed analysis at the stock level. We also hope to provide a theoretical explanation for this phenomenon.

References

Ahn, H.J., Cai, J., Chan, K., Hamao. Y. (2007). "Tick Size Change and Liquidity Provision on the Tokyo Stock Exchange." *Journal of the Japanese and International Economies, 21(2)*, pp. 173-194. https://doi.org/10.1016/j.jjie.2005.10.008

Ahn H.J., Cao, C. Q., Choe, H., (1996). "Tick size, spread, and volume," *Journal of Financial Intermediation*, *5(1)*, pp. 2-22. https://doi.org/10.1006/jfin.1996.0002

Tong, H., (1983). Threshold Models in Non-linear Time Series Analysis. Springer-Verlag New York

Tong, H., Lim K. S., (1980). "Threshold Autoregression, Limit Cycles and Cyclical Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, *42(3)*, pp. 245-268. https://www.jstor.org/stable/2985164

Digital Life

Journal of Digital Life 2025, B-7

Conference Proceedings

Basic Research on the Detection of Dust Mask Wearing Status

Wenyuan Jiang [1], Yuhei Yamamoto [2], Hajime Tachibana [3], Keisuke Nakamoto [3], Kunihiro Katai [3], Daito Yosumi [4]*, Atsuki Yoshida [4] and Aito Yoshikawa [4]

[1]Faculty of Architectural and Environmental Design, Osaka Sangyo University, 3-1-1 Nakagaito, Daito-shi, Osaka 574-8530, Japan
[2]Faculty of Environmental and Urban Engineering, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan
[3]Komaihaltec Inc., 1-19-10 Ueno, Taito-ku, Tokyo 110-8547, Japan
[4]Faculty of Engineering, Osaka Sangyo University, 3-1-1 Nakagaito, Daito-shi, Osaka 574-8530, Japan

1. Introduction

In construction and demolition worksites, large amounts of dust can cause serious occupational diseases such as pneumoconiosis and lung cancer, leading to the 1972 Industrial Safety and Health Act mandating dust mask use. Yet many workers neglect wearing masks due to discomfort or heat, and accidents and lawsuits from dust exposure continue, making proper mask management a challenge. Recent studies have applied ICT-based recognition methods, and we previously proposed a four-class detector (worn correctly, below nose, below chin, no mask). However, in actual factories, workers often turn away from cameras, causing misjudgments. To address this, we add a new category(back of the head) where the mask is not visible to test detection accuracy and explore a more practical approach for workplace blind spots.

2. Methods

In this study, videos taken from construction sites and similar settings are used to determine the mask-wearing status of workers in the video. Furthermore, we devised a classification method that considers multiple patterns of “incomplete wearing” as well as the state where the mask is not visible. In the proposed system, head images of workers are extracted from factory site videos, and VGG19 known for its excellent multi-category classification performance—is used to determine the mask-wearing status. Based on the opinions of factory site supervisors, the mask-wearing status is categorized into five patterns: four commonly observed patterns (worn correctly, worn below nose, worn below chin, and no mask), plus an additional category for cases where the mask-wearing status cannot be determined from the camera’s viewpoint. We labeled this category as “Back of head.”

3. Results

In the experiment, to verify the usefulness of dividing mask-wearing states into five categories, we created training and validation datasets from 60 videos obtained via multiple surveillance cameras at the factory shown in Fig. 1. Since in actual factories all four states of incomplete wearing are relatively rare, we also used the video shown in Fig. 2 to balance the training dataset. The video in Fig. 2 was recorded with five subjects, each performing four mask-wearing states while rotating 360 degrees. First, from the scenes in each video of Fig. 1 where workers appeared, 60 frame images were extracted at equal intervals (a total of 3,600 images). Next, 150 images were randomly selected for each category. If 150 images were insufficient, additional images were obtained from the video shown in Fig. 2.

Fig.1. Research Subject Factory

Fig.2. Video used to Create Part of Training Data

Published: 6 September 2025
* Correspondence: s22k084@ge.osaka-sandai.ac.j

For validation, 10 additional videos of similar factories were used. From scenes with workers, 50 images were extracted at equal intervals, and 10 head images per category were obtained to confirm classification accuracy. During training, 10,000 epochs were set to achieve sufficient accuracy. The final model achieved a loss of 0.003 and an accuracy of 0.97.

The validation results are shown in Table 1.

Table 1. Results of the Generalization Verification Experiment

Prediction / Ground truth	Worn correctly	Worn below nose	Worn below chin	No mask	Back of head
Worn correctly	0	3	6	0	1
Worn below nose	0	2	6	0	2
Worn below nose	1	1	5	0	3
No mask	2	3	2	0	3
Back of head	0	0	0	0	10

From Table 1, it was found that all “Back of the head” images were classified correctly. On the other hand, overall accuracy was low. Many misclassifications occurred as “worn below chin.” In addition, some images were mistakenly classified as “Back of the head.” The possible reasons are as follows:

・Low image resolution
As shown in Fig. 3, some images were of low resolution, making visual judgment difficult even for humans, leading to misclassification.

・Difficulty recognizing side-face features
As shown in Fig. 4, side-face images tend to fail because they simultaneously contain features of both mask wearing and the Back of the head.

・Influence of clothing color
As shown in Fig. 5, in certain head angles, white clothing was often misclassified as a mask

Fig.3. Image with Low Resolution

Fig.4. Image Capturing the Side View of Face

Fig.5. Image Showing Misdetection of Clothing as a Mask

From these observations, while the proposal of a “Back of the head” category is useful, additional categories such as side or oblique angles are also considered necessary.

4. Conclusions

In this research, with the goal of detecting the mask-wearing status of workers in actual factory sites, we introduced the difficult-to-detect category of “back of the head” and confirmed its classification accuracy. The results showed high accuracy for the back-of-head category, but the overall accuracy remained low. The likely reason is that the categories are still insufficient. Therefore, in the future, we plan to consider various additional categories and verify the accuracy.