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

Evaluation of an Environment Classification Method for Optimal Crowd Model Selection in Autonomous Mobile Robot Simulations*

Saki Nakazawa[1] and Yuka Kato[2]

Abstract— **Incorporating crowd models into robot simulators is a common practice in the field of autonomous mobile robot navigation research. Although crowd models developed in the field of crowd simulation are frequently used, there is no universal model that can be used for all scenarios. Therefore, it is essential to select an appropriate crowd model for the specific simulation environment. From this perspective, until now, we have been studying a method for selecting an appropriate crowd model for each category. This method observes the movement trajectories of pedestrians in the environment to be simulated, and classifies the environments into multiple categories based on the observation results. In this process, feature images are generated by superimposing the observation results on a time axis. The latent variables in the feature images are then extracted using an autoencoder, and the extraction results are clustered. However, the impact of overlapping time intervals (temporal granularity) and different extraction methods of latent variables, which are important in generating feature images, has not yet been clarified. In this paper, we assess their influence on category classification accuracy. Based on the obtained classification results, we also develop scenarios for selecting a crowd model for each category.**

I. INTRODUCTION

Incorporating crowd models into robot simulators is a common practice in the field of autonomous mobile robot navigation research [1]. Here, crowd models developed in the field of crowd simulation are frequently used, such as Social Force Model (SFM) [2] and Optical Reciprocal Collision Avoidance (ORCA) [3]. However, there is no general-purpose model that can handle all scenarios (environmental geometry, crowd movement trends, human attributes, etc.), pointing out the necessity of selecting an appropriate crowd model for each simulation environment [4], [5]. Since the crowd model selection affects navigation performance, how to select the appropriate model for the environment has become an important research question.

With this background, we have been studying a method to select an appropriate crowd model for each category by observing pedestrian movement trajectories in the environment to be simulated and classifying the environment into multiple categories based on the observation results [6]. Specifically, (i) feature images are generated by superimposing the observation results on a time axis; (ii) latent variables in the generated images are extracted using an autoencoder; and (iii) the variables are clustered for categorization. The effectiveness of the method has been verified by regenerating feature images from the classification results and visually comparing their accuracy. However, the impact of differences in temporal granularity (the time interval at which observation results are superimposed on the time axis) and latent variable extraction methods for clustering has not been clarified. The proposed method is characterized by its representation of a series of data (movement trajectories of crowds) that changes with time as a single image superimposed on a time axis. There would be an appropriate temporal granularity for classification. In addition, although one-dimensional arrays have been used as input to the autoencoder, it is considered that more suitable methods exist for extracting image features.

Based on the above, this paper conducts the following two evaluations: (i) an evaluation in which the temporal granularity is changed to 1 hour and 10 minutes; and (ii) an evaluation in which the latent variable extraction method is changed to Autoencoder (AE), Convolutional Autoencoder (CAE), and Variational Autoencoder (VAE). As evaluation metrics, the inter-cluster separation, the inner-cluster compactness (consistency), and the Silhouette score are used, in addition to the visual evaluation (subjective evaluation) previously employed. The contributions of this paper are:

- Evaluate the impact of differences in temporal granularity and latent variable extraction methods on classification accuracy in the proposed method.
- Develop possible scenarios for selecting a crowd model for each category based on the classification results.

II. RELATED WORK

There are a variety of crowd models incorporated into robot simulators, which are often categorized into three types: force-based models, velocity object (VO) based models, and visual-based models. Force-based models assume virtual repulsion from the environment and surrounding agents and use the equation of motion to determine the behavior of each agent, and SFM [2] is a typical example. VO-based models determine the behavior by sequentially predicting the future movements of surrounding agents at short intervals based on the agents' velocities, and ORCA [3] is a typical example. Vision-based models determine the behavior based on visual information such as camera images (e.g., Vision Based Navigation (VBN) [7]). Each model has its advantages and disadvantages, and ORCA-based methods are often used in robot simulators. However, the validity of applying a single model regardless of the scenario is not clear. Recently, various methods to generate movement

*This work was partially supported by JSPS KAKENHI (23K11087, 23K11073), the Telecommunications Advancement Foundation, the Cooperative Research Project Program of RIEC, Tohoku University, and Research Grant of TWCU.

[1]Saki Nakazawa was with Graduate School of Science, Tokyo Woman's Christian University, Tokyo, Japan. d23m202@cis.twcu.ac.jp

[2]Yuka Kato is with Division of Mathematical Sciences, Tokyo Woman's Christian University, Tokyo, Japan. yuka@cis.twcu.ac.jp

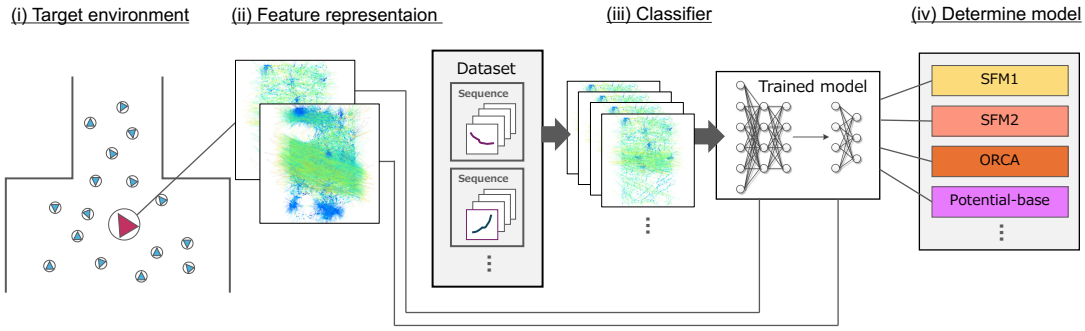

Fig. 1: Category classification process: (i) based on the movement trajectory of the crowd observed by the environmental sensors; (ii) generate feature representations (i.e., feature vectors) that characterize the spatio-temporal data; and (iii) using a pre-trained classifier generated in advance by machine learning algorithms with datasets; (iv) determine the category of the target environment [6].

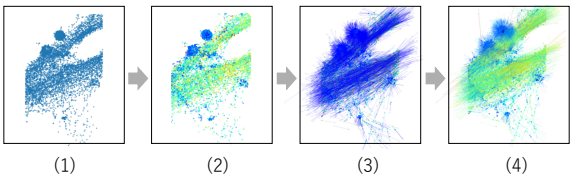

Fig. 2: Feature image generation procedure. Colors and arrow directions represent characteristics of movement trends.

trajectories for each agent using machine learning algorithms have been proposed [8], [9], but these methods require model training for each scenario and are often not easy to use.

For this reason, several methods have been proposed to classify the movement tendency of crowd into several categories based on observations and use the classification results for crowd model selection [10], [11]. We have also proposed a method that generates a training dataset for machine learning by classifying the environment based on pedestrian movement trends [12]. These studies have mainly used approaches that classify the visualization results of congestion and human movement trends using machine learning algorithms. However, it is not clear what kind of features should be used as factors to characterize the time-space when targeting simulators for robotics. The literature [6] focuses on this point and proposes a method using a velocity vector.

III. METHOD

A. Overview

First, we present an overview of the environment classification method based on [6]. The category classification process is shown in Fig. 1. Based on the movement trajectory observed by the environmental sensors, the method generates feature representations that characterize the spatio-temporal data, and determines the category of the target environment by using a pre-trained classifier generated by machine learning algorithms with datasets. After that, the crowd model to be used is determined according to the category.

B. Problem Formulation

To train the classifier, we use a dataset of observed and recorded pedestrian movement trajectories in various environments. It is assumed that the dataset consists of timestamps, pedestrian IDs (an ID that identifies a walking trajectory), and the position coordinates of the pedestrians (x, y) at the measurement time. The training data is constructed by extracting multiple subsets from the dataset.

Specifically, we extract the sets of data contained in certain areas and certain durations from the dataset and construct N subsets $\mathcal{D}^{[1]}, \mathcal{D}^{[2]}, \cdots, \mathcal{D}^{[N]}$, and generate the feature representation of $\mathcal{D}^{[i]}$ as $\boldsymbol{x}_i = f(\boldsymbol{s}_i)$ $(i = 1, 2, ..., N)$ using a certain function f, where $\boldsymbol{s}_i \in \mathcal{D}^{[i]}$ and $\boldsymbol{x}_i \in \mathbb{R}^d$. This set of feature representations $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\} = \{\boldsymbol{x}_i\}_{i=1}^N$ is used to construct the classifier. The number of clusters is K, and the elements of $\mathcal{X}$ are classified into one of the clusters $\mathcal{Y} = \{C_1, C_2, ..., C_K\}$ based on some degree of similarity. The classification criteria obtained in this process are used as the categorical classifier. For categorization, we construct $\boldsymbol{s}^*$ by observing and accumulating the position coordinates (x_j, y_j) of the pedestrian j at a certain time t in the simulated environment, and determine the appropriate category $y^* \in \mathcal{Y}$ by using $\boldsymbol{x}^* = f(\boldsymbol{s}^*)$.

C. Feature Image Generation

This method uses an image as the feature representation $\boldsymbol{x}_i$ of the i-th subset $\mathcal{D}^{[i]}$. Specifically, the feature vector is generated by superimposing images representing the position coordinates of the pedestrians, movement speeds, and directions of movement at time intervals that constitute the subset $\mathcal{D}^{[i]}$. Translucent images are used to represent the shading (density) caused by the superimposition, and color mapping is used to change the color for visualizing the magnitude of the values. The function $f(\boldsymbol{s}_i)$ is expressed as follows:

1) Draw the coordinates (x, y) of all pedestrians in $\mathcal{D}^{[i]}$ as a scatter plot using translucent images.
2) Change the color of the points, making them more blue/yellow when moving slowly/quickly.
3) Calculate the velocity vector for each pedestrian and draw it as a translucent arrow overlaid on the scatter plots. The length and direction represent the magnitude of the speed and movement direction.
4) Change the color of the arrows, making them more blue/yellow when moving slowly/quickly.

The procedure of feature image generation is shown in Fig. 2.

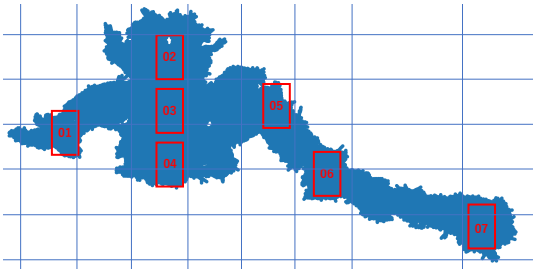

Fig. 3: Overall shape of the shopping mall (scatter plots of the coordinates where pedestrians are present) and the areas corresponding to the target environments (01 $\sim$ 07) [6].

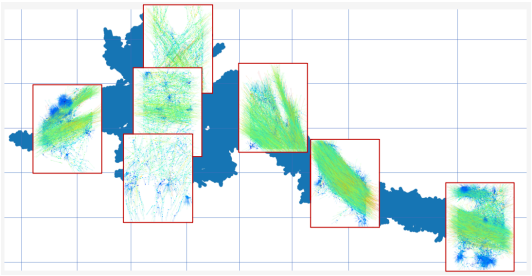

Fig. 4: Example of extracted feature image for each area.

D. Category Classification

The category classifier is constructed by using the generated $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$. This method extracts latent variables from the generated images using three types of training models: Autoencoder (AE), Convolutional Autoencoder (CAE), and Variational Autoencoder(VAE). Then, the extraction results are clustered to classify categories.

In the literature [6], only AE is used as the training model. In this paper, CAE and VAE are added to the model to evaluate the impact of different latent variable extraction methods on classification accuracy. CAE has the feature of preserving the spatial structure of the input data by merging CNN and AE; VAE has the potential to extract data characteristics more flexibly by explicitly modeling the distribution of data in the latent space. Since this method aims to classify artificially generated abstract images (e.g., Fig. 2), we add CAE, which preserves the spatial structure, and VAE, which allows for flexible feature extraction.

IV. EXPERIMENTAL SETTINGS

To evaluate the impact of differences in the temporal granularity of feature image generation and the latent variable extraction methods on classification accuracy, we conducted an evaluation experiment using a pedestrian movement trajectory dataset. Here, we actually categorized the simulation environments based on the classification method described in Section III, and compared the results based on evaluation metrics. The comparisons were made by changing the temporal granularity to 1 hour and 10 minutes, and by changing the latent variable extraction method to AE, CAE, and VAE. The experimental settings are described in detail below.

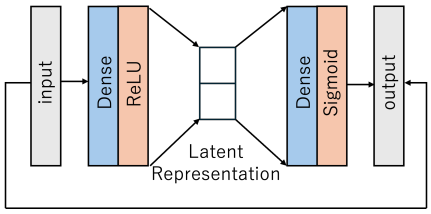

Fig. 5: Autoencoder (AE) Model.

A. Dataset

We used the datasets from the ATC shopping center [13] collected at a shopping mall in Japan (`https://dil.atr.jp/crest2010_HRI/ATC_dataset/`). The entries of the dataset are as follows:

- Date: Wed. and Sun. from 10/24/2012 to 11/29/2013
- Time: 9:00 – 21:00
- Items: time [ms], person ID, posision (x, y, z) [mm], velocity [mm/s], motion angle [rad], facing angle [rad]

B. Feature Image Generation

The subsets were constructed by extracting 7 areas with different characteristics from the whole dataset with reference to the area of a small exhibition room in a museum ($5,600$ mm $\times$ $9,500$ mm). The overall shape of the shopping mall and the areas corresponding to the target environments are shown in Fig. 3 (numbered from 01 to 07). The reason for selecting areas with different characteristics is to ensure that the diversity of crowd behavior is reflected in the classifier. The target dataset contains a variety of spatial structures and crowd behavior patterns, and selecting areas with different characteristics is expected to cover a wide range of scenarios. The results of extracting representative feature images are shown in Fig. 4. Each area has the following features.

- Area 01: Near the entrance and exit to the parking lot, heavy pedestrian traffic, and few pedestrians staying.
- Area 02: Near the entrance and exit to the train station, and low pedestrian traffic.
- Area 03 and 04: Many pedestrians staying in the square.
- Area 05, 06, and 07: Passage from the station to the square, heavy pedestrian traffic, and few staying.

Fig. 4 indicates that each area has visually different features, but note that these features change from time to time, and their images are only examples.

During image generation, the pixel values of the images were normalized by scaling them from 0 to 1, and all images were resized to 128×128 pixels for the same resolution. The number of feature images generated was 7,569 with a temporal granularity of 1 hour and 6,083 with a temporal granularity of 10 minutes.

C. Classifier Construction

The dimension of the input image was set to $d = 128 \times 128 \times 3$ (RGB image), which was used as input to AE, CAE, and VAE with a 1,000-dimensional feature vector as a latent representation (for all models). As for the clustering method, we used the k-means clustering ($K = 5$) similar to [6].

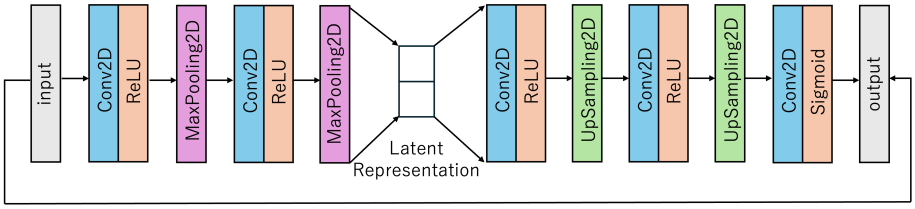

Fig. 6: Convolutional Autoencoder (CAE) Model.

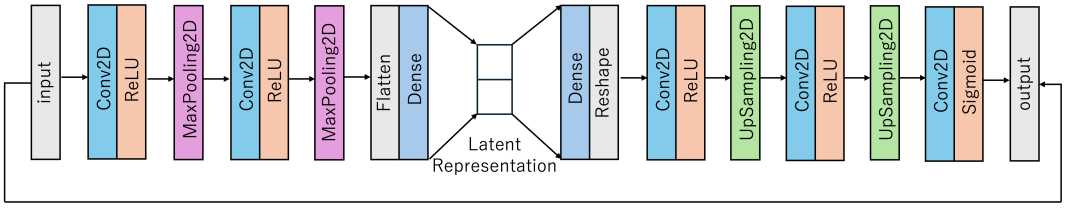

Fig. 7: Variational Autoencoder (VAE) Model.

For the latent variable extraction methods used in the experiments, the structure of the AE model is shown in Fig. 5, that of the CAE model in Fig. 6, and that of the VAE model in Fig. 7. In AE, RGB image data was flattened into a one-dimensional vector as input, Adam was used as the optimizer, and Binary Cross Entropy was used as a loss function. In CAE, different loss functions are used to minimize the reconstruction error of the input data, and in this experiment, Binary Cross Entropy (BC), Mean Absolute Error (MAE), and Mean Squared Error (MSE).

D. Evaluation Metrics

As evaluation metrics, we used the inter-cluster separation, the inner-cluster compactness, and the Silhouette score, in addition to the visual evaluation (subjective evaluation).

1) Inter-cluster separation: To evaluate the degree of clarity of separation between categories, we used the inter-cluster separation. It calculates the distance between centroids (Euclidean distance) for each cluster; the larger the value, the greater the differences between clusters. For each cluster C_i, calculate the centroid μ_i and obtain the degree of separation as the average of $D_{ij} = \| \mu_i - \mu_j \|$.

2) Inner-cluster compactness: To evaluate the similarity within a category, we used the inner-cluster compactness (consistency score). It calculates the average distance between the data in a cluster; the smaller the value, the higher the similarity within a cluster. The inner-cluster compactness IC_i for cluster C_i is

$$IC_i = \frac{2}{|C_i|(|C_i|-1)} \sum_{x,y \subset C_i, x \neq y} \| x - y \| \qquad (1)$$

where $|C_i|$ is the number of data in cluster C_i and x, y is the data values in cluster C_i. The score is calculated as the average of these values for all clusters.

3) Silhouette score: To evaluate the overall quality of the clustering results, we used the Silhouette score. This score is calculated as the average of the Silhouette coefficients for each data, where the coefficients are calculated as an overall assessment of how well a given data fits into its own cluster and how different that data is from other clusters. The calculation procedure is as follows:

1) For each data i, calculate the average distance $a(i)$ from all other data in its cluster.
2) Calculate the average distance $b(i)$ from other clusters that do not contain data i.
3) Calculate the Silhouette coefficient $s(i)$ as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (2)$$

4) The Silhouette score is the average of $s(i)$.

V. EXPERIMENTAL RESULT

A. Subjective Evaulation

As a subjective evaluation, we obtained decoded images of the clustering results, and visually compared five randomly selected images for each category side by side in a row. The clustering results for each latent variable extraction method at a temporal granularity of 1 hour are shown in Fig. 8, and those of 10 minutes are shown in Fig. 9. Here, images in the same cluster are placed horizontally side by side.

Overall, the separation between clusters is clearer in the 1-hour case than in the 10-minute case. As for the latent variable extraction methods, clustering using CAE (BC), CAE (MAE), and VAE showed relatively high consistency, with little mixing of different categories. In particular, VAE shows high consistency and clear cluster boundaries, with excellent clustering performance in the 1-hour case.

B. Objective Evaluation

As an objective evaluation, we obtained the inter-cluster separation, the inner-cluster compactness, and the Silhouette score. The experimental results for temporal granularity and latent variable extraction method are shown in Table I.

The inner-cluster compactness is the lowest for AE in both 1-hour and 10-minutes cases, while the inter-cluster separation and the Silhouette score are the highest for VAE.

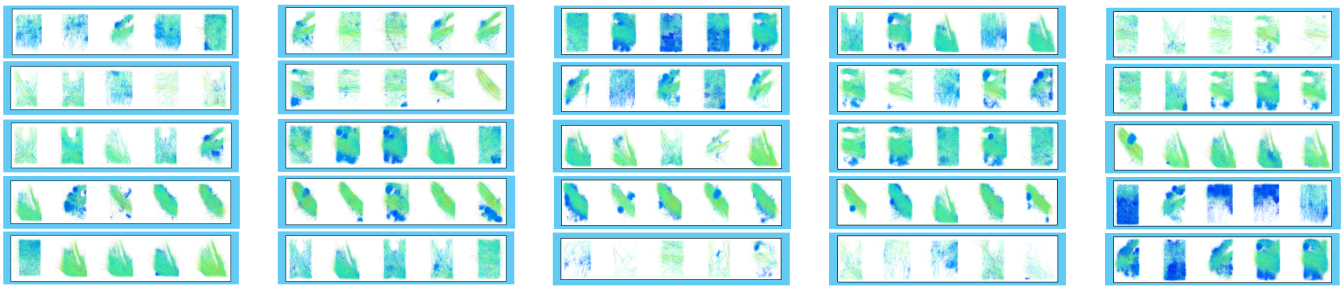

Fig. 8: Visualization of clustering results for each latent variable extraction method at a temporal granularity of 1 hour.

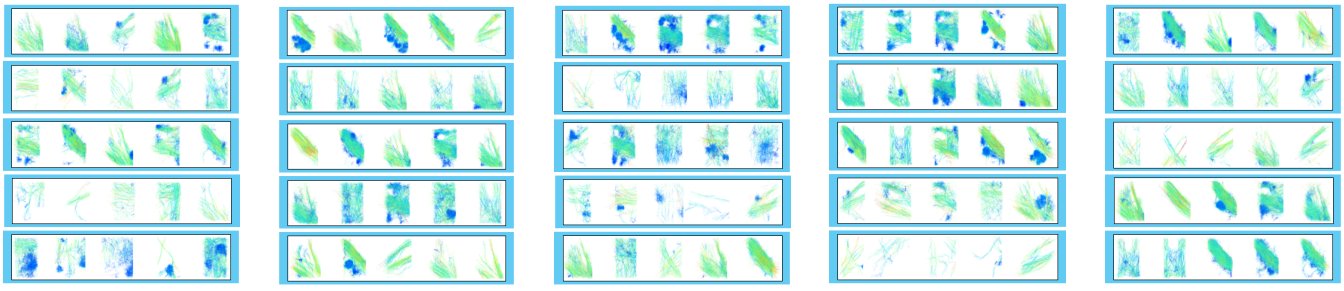

Fig. 9: Visualization of clustering results for each latent variable extraction method at a temporal granularity of 10 minutes.

TABLE I: Experimental results of objective evaluation for temporal granularity and latent variable extraction method.

Temporary granularity	Extraction model	Inter-cluster separation	Inner-cluster compactness	Silhouette score
1 hour	AE	0.79	**0.88**	0.24
	CAE (BC)	1.00	8.18	0.27
	CAE (MAE)	14.04	13.64	0.34
	CAE (MSE)	9.93	7.88	0.35
	VAE	**44.24**	11.99	**0.58**
10 minutes	AE	0.64	**0.97**	0.16
	CAE (BC)	9.92	9.48	0.20
	CAE (MAE)	13.24	13.27	0.30
	CAE (MSE)	8.19	7.32	0.27
	VAE	**31.05**	7.11	**0.58**

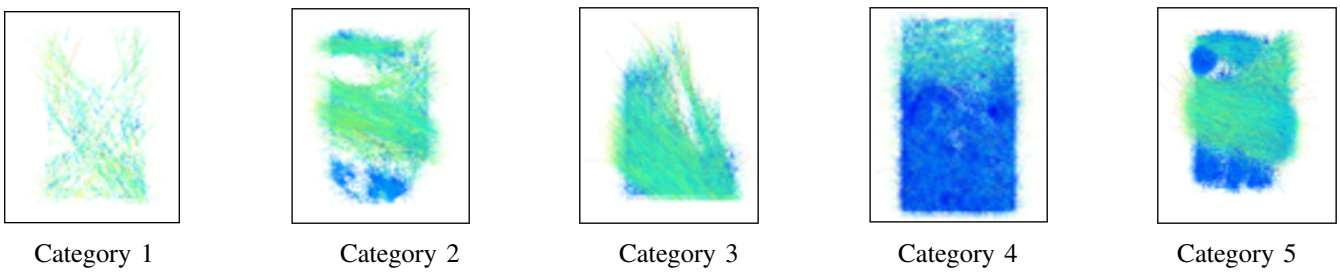

Fig. 10: Representative images extracted from each category.

Combined with the results of the subjective evaluation, VAE is considered superior to the other methods. For temporal granularity, the overall evaluation value is higher for the 1-hour case, regardless of the latent variable extraction method.

C. Discussion

In this section, we discuss the impact of differences in temporal granularity and latent variable extraction methods.

Regarding the effect of temporal granularity, VAE showed stable performance for both 1 hour and 10 minutes, whereas the performance of the other methods varied with them. In particular, CAE (MAE) and CAE (MSE) tend to perform poorly at the 10-minutes case, and have problems with robustness to short-time data variations.

For latent variable extraction methods, VAE significantly outperformed the other methods on all evaluation metrics.

This suggests that VAE better captures the structure of the data in the latent space and has a superior ability to clearly separate each cluster. In particular, VAE shows high values of the inter-cluster separation and the Silhouette score, indicating that it generates clear boundaries between clusters. For that, less mixing of different categories and more consistent clustering are realized. On the other hand, the performance of CAE varied greatly depending on the loss functions. CAE (BC) and CAE (MAE) showed high inner-cluster compactness and a moderate degree of inter-cluster separation, with MAE in particular showing less inner-cluster variation. This is because MAE is robust to outliers and has the effect of bringing the data in a cluster closer together. However, CAE (MSE) showed slightly poorer performance than the other CAE methods, especially at a temporal granularity of 10 minutes. MSE is sensitive to outliers, which may increase the variability of the data. AE has lower clustering performance than the others, with lower inter-cluster separation and Silhouette score, indicating that it is difficult to form consistent clusters.

These results indicate that it is important to select a suitable temporal granularity and variable extraction method.

VI. APPLICATION

Based on the experimental results, we developed a set of assumed scenarios for selecting a crowd model for each category. Here, we used the clustering result for temporal granularity of 1 hour and VAE, which are considered to have the best clustering performance. We first extract typical representative examples from the subset classified into each category, and consider these characteristics of the area and time period as the features of each category. Then, we map the categories to the crowd models by associating the characteristics with the crowd models.

First, we extracted a feature image representing each category from the experimental results. The images are shown in Fig. 10. The following summarizes the area, time periods, and characteristics from which each image was extracted:

- Category 1: Area 02 (15:00 – 16:00), low pedestrian traffic near the entrance and exit to the train station. This area is characterized by moderate traffic volume and few pedestrian staying, making for a smooth movement.
- Category 2: Area 07 (14:00 – 15:00), a passage from the station to the square. This area is characterized by high traffic volume and few pedestrian staying.
- Category 3: Area 05 (14:00 – 15:00), a passage from the station to the square. This area is characterized by high traffic volume and few pedestrian staying.
- Category 4: Area 03 (15:00 – 16:00), open square. This area is characterized by many staying in the square.
- Category 5: Area 07 (15:00 – 16:00), heavy pedestrian traffic in the passage from the station to the square. This area is characterized by holidays when shopping malls are crowded and people stay inside the stores.

We considered these characteristics as category features and summarized them in the following four scenarios.

- Scenario 1: Low density, few stay, moderate speed
- Scenario 2: Moderate density and speed, partial stay
- Scenario 3: High density, stay, low speed
- Scenario 4: High density, stay, moderate speed

These can be associated with crowd models by using model selection methods appropriate for a scenario [14].

VII. CONCLUSIONS

In this paper, we evaluated the impact of different temporal granularities (1 hour and 10 minutes) and latent variable extraction methods (AE, CAE, VAE) on classification accuracy. The results of the study can be summarized as follows:

- Effect of temporal granularity: classification accuracy was higher for the 1-hour case than for 10-miniutes case. In particular, VAE showed superior performance.
- Effect of feature extraction method: VAE showed the highest performance. The ability to capture complex feature space may contribute to the performance.

In the future, we will determine a more appropriate temporal granularity and design evaluation metrics considering crowd model selection.

REFERENCES

[1] K. Amano, A. Komori, S. Nakazawa, and Y. Kato, "Impact of environment on navigation performance for autonomous mobile robots in crowds," in *Proc. IEEE/SICE International Symposium on System Integration (SII 2023)*, 2023, pp. 794–799.

[2] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[3] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research*, 2011, pp. 3–19.

[4] T. Fraichard and V. Levesy, "From crowd simulation to robot navigation in crowds," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 711–716, 2020.

[5] M. Tanaka and Y. Kato, "Evaluation of crowd models under various environments for robot navigation simulator," in *Proc. IEEE/SICE International Symposium on System Integration (SII 2024)*, 2024, pp. 794–799.

[6] S. Nakazawa and Y. Kato, "Environment classification method using autoencoder to select appropriate crowd model for robot simulation," in *Proc. IEEE International Conference on Autonomous Science and Engineering (CASE 2024)*, 2024.

[7] T. Dutra, R. Marques, *et al.*, "Gradient-based steering for vision-based crowd simulation algorithms," *Computer Graphics Forum*, vol. 36, no. 2, pp. 337–348, 2017.

[8] N. Bisagno, B. Zhang, and N. Conci, "Group lstm: Group trajectory prediction in crowded scenarios," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[9] J. Amirian, W. van Toll, J.-B. Hayet, and J. Pettré, "Data-driven crowd simulation with generative adversarial networks," in *Proc. of the International Conference on Computer Animation and Social Agents (CASA)*, 2019, pp. 7–10.

[10] I. Karamouzas, N. Sohre, R. Hu, and S. J. Guy, "Crowd space: A predictive crowd analysis technique," *ACM Trans. on Graphics*, vol. 37, no. 6, pp. 186.1–186.14, 2018.

[11] J. Carvalho, M. Marques, and P. Costeira, "Understanding people flow in transportation hubs," *IEEE Trans. on Intell. Transport. Sys.*, vol. 19, no. 10, pp. 3282–3291, 2018.

[12] R. Akabane and Y. Kato, "Pedestrian trajectory prediction based on transfer learning for human-following mobile robots," *IEEE Access*, vol. 9, pp. 126 172–126 185, 2021.

[13] D. Brscic, T. Kanda, T. Ikeda, and T. Miyashita, "Person position and body direction tracking in large public spaces using 3D range sensors," *IEEE Trans. on Human-Machine Systems*, vol. 43, pp. 522–534, 2013.

[14] R. Nishida and Y. Kato, "A crowd model evaluation method for autonomous mobile robot simulator," in *International Conference on Human System Interaction (HSI 2024)*, 2024, pp. 1–6.