A Multi-Scenario Crowd Data Synthesis Based On Building Information Modeling

Huang H.[1], Gao G.[1,2], Ke Z.[1], Peng C.[1], Gu M. [1,2]
[1]School of Software, Tsinghua University, Beijing, China, [2]Beijing National Research Center for Information Science and Technology(BNRist), Tsinghua University, Beijing, China
gaoge@tsinghua.edu.cn

Abstract. Deep learning methods have proven to be effective in the field of crowd analysis recently. Nonetheless, the performance of deep learning models is affected by the inadequacy of training datasets. Because of policy implications and privacy restrictions, crowd data is commonly difficult to access. In order to overcome the difficulty of insufficient dataset, the previous work used to synthesize labelled crowd data in outdoor scenes and virtual games. However, these methods perform data synthesis with limited environmental information and inflexible crowd rules, usually in unauthentic environment. In this paper, a tool for synthesizing crowd data in BIM models with multiple scenes is proposed. This tool can make full use of the comprehensive information of real-world buildings, and conduct crowd simulations by setting behavior rules. The synthesized dataset is used for data augmentation for crowd analysis problems and the experimental results clearly confirm the effectiveness of the tool.

1. Introduction

In recent years, deep learning has become a technology area of great interest. A large number of models based on deep learning algorithms have emerged in the field of computer vision. These deep learning models help solve some crowd analysis problems that are difficult for traditional methods, including crowd detection (Liu et al.,2019), trajectory prediction (Huang et al.,2019), pose estimation (Kocabas et al.,2019), crowd counting (Liu et al.,2019) and so on. However, the efficient training of deep learning models relies on a large amount of labeled data. The production of datasets requires high-quality raw data and accurate labeling information, which is time-consuming and laborious. Data is also increasingly difficult to access because of policy implications and privacy restrictions. Currently, the performance of models in crowd trajectory prediction, crowd counting and other tasks is insufficient partly due to the shortage of training datasets. In order to improve the performance of crowd analysis models, researchers tried to develop some alternative methods to acquire datasets. Recently, synthetic datasets are widely used in deep learning field to improve model performance.

For crowd data synthesis, some researchers have developed data collectors and annotators to generate large-scale crowd data from games. And other researchers consider using graphics simulators to generate crowd data in real scenes. Some datasets (Richter et al.,2016; Richter et al.,2017; Wang et al.,2019) collect virtual scenes through GTA5 (a computer game) and then automate the annotation. Specifically, Richter et al(2016) reconstructs the communication between the game and the graphics hardware and successfully completes semantic annotation of the game data while the game code is not available. Richter et al(2017) proposes a new middleware construct that receives render commands from the game. Then, the data obtained by the middleware is annotated by the system in real time, and finally the usable dataset is generated. Wang et al(2019) develops a data collector and annotator that could generate synthetic labeled crowd scenes. Moreover, the method (Wang et al.,2019) is not limited by the characteristics of GTA5 (population limitation) through a split-and-splice strategy, which enables it to produce complex crowded scenarios. A mixed reality dataset was proposed by Cheung et al(2018), which combined background images of the real world with synthetic

https://doi.org/10.7146/aul.455.c223

pedestrians for pedestrian detection. Chai et al(2020) uses neural network model to successfully generate a continuous crowd video with diverse crowd behaviors. A crowd data synthesis tool based on real image scene is proposed by Khadka et al(2020). Based on graphics tools, the real image scenes can be used to generate data sets for a variety of computer vision-related problems. The results of quantitative crowd analysis confirm the success of using synthetic data to train and test networks.

However, these datasets focus on graphical generation and ignore the impact of the environment and crowd movement rules. In previous works, crowd data is generated in empty outdoor scenes or in virtual games, wherefore the source of pedestrian environmental information is either the virtual world, or a few images with limited information. Moreover, the rules of crowd behavior are not static, but change according to environmental information and crowd status. Since the game code is not available, the crowd movement rules are hard-coded and cannot be changed according to the environment. In addition, due to the incomplete information, the previous works only synthesize the crowd from a single and partial perspective, which leads to poor generality and reusability.

In this paper, therefore, a tool for generating crowd data in real buildings with multiple scenes is proposed. By utilizing real world building information models (Eastman et al.,2008) rich in geometric and semantic information, our tool bridges the gap between the real world and the virtual environment. Since there are industrial standards for Building Information Modeling software, we can obtain a large amount of realistic data in an efficient way, so as to greatly increase the credibility of the generated data. Besides, Multi-scenario crowd simulation allows crowd behavior rules to be set by users. The simulation results can be combined with building information models to render realistic crowd data and provide a credible ground reality and annotation. In addition, datasets synthesized with our tools has complete crowd behavior data and environmental information and can be directly used for various complex computer vision tasks, such as trajectory prediction, crowd counting, and other indoor tasks.

To the best of our knowledge, our work is the first to generate indoor crowd data on the basis of real-world buildings. Several key contributions of our work are:

- The novel tool extracts information from industry-standard realistic building models that can be obtained in bulk and with authenticity.
- According to the simulation results, the tool can use graphical rendering to synthesize datasets of various crowd analysis problems, which are automatically labeled.
- The data generated by the tool can improve the results of existing crowd analysis tasks, and related new crowd analysis problems can also be proposed.

2. Methodology

In this section, the proposed tool algorithm flow and implementation method are described. BIM model information is firstly extracted, followed by the completion of multi-agent population simulation. Then, the indoor crowd data is synthesized and stored. Figure 1 shows the architecture of the proposed tool.

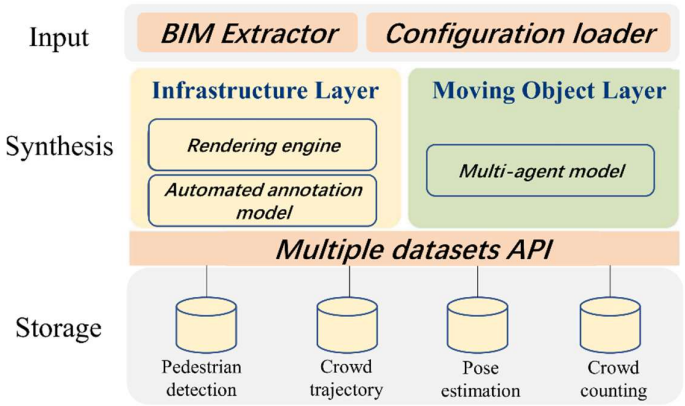

Figure 1: System architecture

2.1 Information extraction from BIM models

Building information modelling is the 3D CAD modelling technology for various buildings widely used in Architecture, Engineering, and Construction (AEC) (Eastman et al.,2008) industries. It stores real-world buildings in digital form, including geometric information and semantic information. The Industry Foundation Classes (IFC) data format is the internationally accepted specification for BIM, and thus used as the building model format in this study.

We mainly extract the geometric information of the building and the semantic information of each component (such as doors, walls, rooms and spaces) from the BIM model to form our navigation map, as shown in Figure 2. The geometric information helps us to obtain the boundaries of spaces and products in buildings. And the semantic information can help us get the passability and risk coefficient of different areas. We specifically discuss BIM information extraction process as follows.

Geometric information. As a part of a navigational map, any building component with a shape can be a pedestrian-accessible space (e.g., stair) or a pedestrian-inaccessible space (e.g., wall). Through the geometry information provided in BIM, we can calculate the bounding box of the 3d building components and place them in local placement.

Semantic information. In traditional maps, semantic information is generally ignored. However, in the process of crowd simulation, the state of component (e.g., locked door) or space (e.g., private room) will have a great influence on the simulation results. And some of the building components may affect the mobility of people.

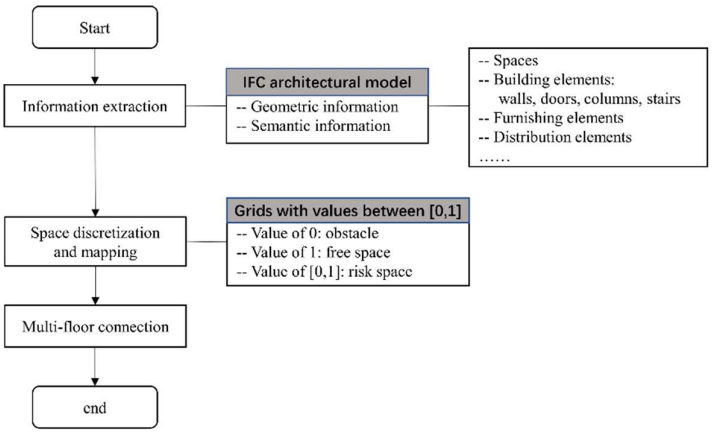

Figure 2: General data flow for BIM extraction

Once BIM information extraction is completed, the building indoor map with rich semantic information is constructed. Figure 3 shows the overall appearance of a BIM model and Figure 4 shows the internal perspective of this BIM model.

Figure 3: The facade of a BIM model

Figure 4: The internal perspectives of a BIM model

2.2 Multi-scenario crowd simulation

Multi-agent methods (Sharma et al.,2018) are often used to simulate the complex behavior of various groups. The rules of multi-agents can be defined by users, which is proven to be reliable in crowd simulation. In this section, we use the multi-agent model to simulate individuals with independent consciousness. Moreover, we set up different multi-agent targets to meet the needs of different scenarios. In crowd simulation, human attributes are assigned to the agent, including age, moving speed, occupied space and random fall factors. In addition, different behavior rules are designed according to different scenes, so that agents can properly interact with other agents and the environment. Figure 5 shows the flow of multi-agent simulation process.

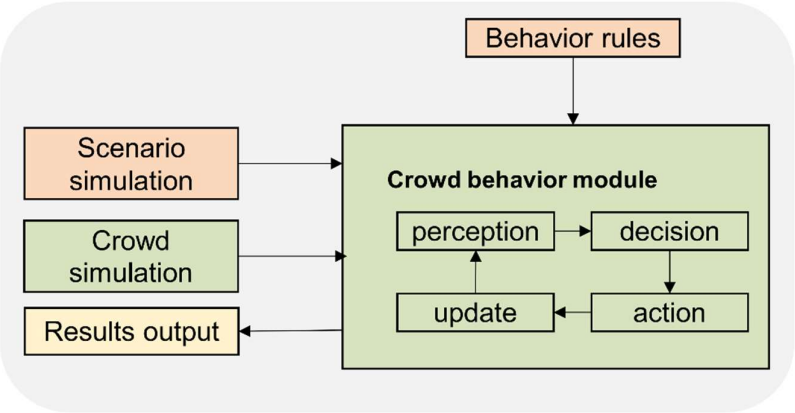

Figure 5: Multi-agent simulation process

For this simulation algorithm, some configuration parameters such as population distribution, age ratio and simulation scenarios are inputted by users. After the configuration file is set up, the tool could perform crowd simulation in normal and emergency situations. The scenario of the normal situation is to generate the crowd data in the daily operation process of the building. It has no specific purpose and is suitable for general computer vision tasks. The scenario of the emergency situation is to generate the crowd data in the condition of a critical event.

After completing the simulation, the tool will get a series of continuous agent coordinates and state results. The result of an agent is represented by a quad vector:

$$a = \{(X^t, Y^t, Z^t, S^t | t \in T)\} \quad (1)$$

where (X^t, Y^t, Z^t, S^t) is the agent's position and current state at time *t*. In this quad, X^t and Y^t are plane coordinates, expressed in absolute coordinates. Since the elevation of the agent in the building is consistent with that of the floor on which the agent is located, the height coordinate Z^t directly adopts the elevation of the floor. S^t is the current state of the agent, which is very important for subsequent data synthesis. Table 1 lists all possible states of the agent.

Table 1: The state category of the agent.

S	State of the agent
0	Resting
1	Walking
2	Falling
3	Running
4	Talking

The simulation results of all the agents are denoted by

$$A = \{a_i | i = 1,2, \dots, N\} \quad (2)$$

2.3 Data generation

Unlike generating datasets from game data or simple pictures, environment information and crowd information are fully accessible and computable in our tool. Comprehensive information makes it possible to annotate datasets accurately and automatically.

Specifically, the simulation results will be rendered in BIM model scenes as crowd information. Whereafter, the tool will record rendering results from multiple indoor perspectives to produce video or images. According to different dataset requirements, the tool extracts, combines and calculates the corresponding information from the crowd simulation results, and automatically complete the annotation of the dataset. The process of data generation is shown in Figure 6.

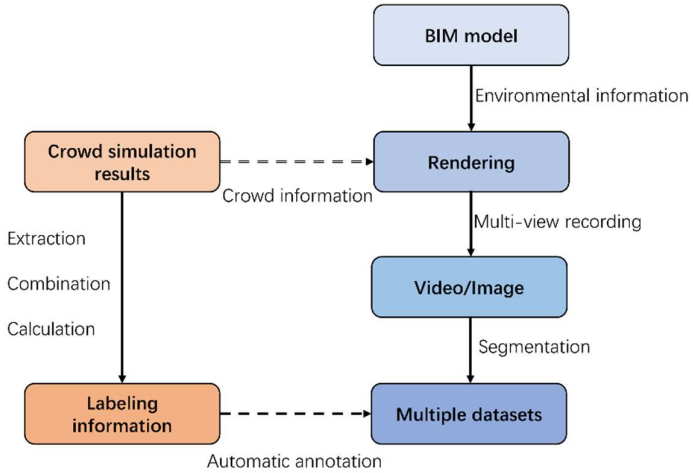

Figure 6: The process of data generation

This tool builds multiple datasets for the following four tasks: 1) pedestrian detection, 2) trajectory prediction, 3) pose estimation, and 4) crowd counting.

Pedestrian detection. Crowd detection needs to collect datasets with human body bounding box information. The tool calculates the bounding box at different times according to the size and state of the agent from the simulation results, and then automatically labels the bounding box of the coordinate where the agent is located.

Trajectory prediction. Trajectory prediction requires pedestrian trajectory data within a period of time, which is expressed as:

$$traj = \{(X^t, Y^t | t \in T)\} \tag{3}$$

As we can see, the crowd simulation result is almost the same as the trajectory dataset. In the tool, the trajectory of the crowd can be easily obtained by extracting the trails of moving people in the field of view.

Pose estimation. In the pose estimation task, the joint positions of the human body need to be used. Since the crowd is rendered in 3D by the graphics renderer, the tool can directly get the joint positions of the crowd from graphical engine. Therefore, the acquired joint sites can be labeled into the video or images.

Crowd counting. The head coordinates of the crowd can be obtained from the rendering results, and the tool will automatically label the crowd in the field of view to get a crowd counting datasets.

(a) Crowd detection (b) Trajectory prediction

(c) Pose estimation (d) Crowd counting

Figure 7: Datasets generated by our tool

3. Evaluation

Trajectory prediction is a popular task in crowd analysis. In this section, we conducted experiments on two relevant datasets: ETH (Pellegrini et al.,2010) and UCY (Lerner et al.,2007). These datasets contain different types of scenes: Zara1, Zara2, Univ, Eth, and Hotel. We chose Social-BiGAT (Kosaraju et al.,2019) and Sophie (Sadeghian et al.,2019) as the experimental network architectures to confirm the improvement on trajectory prediction with our synthetic data. Both networks use the combination of scene information and social information for trajectory prediction.

In the model test phase, the models observe the crowd trajectory in the past 3.2 seconds, and then predict the trajectory in the next 4.8 seconds. Two metrics will be used to evaluate the performance of the experiment: average displacement error (ADE) (Pellegrini et al.,2009), and final displacement error (FDE) (Alahi et al.,2016). Both are used to evaluate the average distance between the true trajectory and the predicted result.

In order to prove the effectiveness of synthetic data, we pre-trained the models with synthetic data, and compared with the original performance. The results of the experiments are shown in Table 2.

As expected we see that both the networks trained with synthetic data outperform those trained without synthetic data. It is proved that our synthetic data can improve the performance of crowd trajectory prediction, regardless of the network structure.

Table 2: ADE/FDE for Social-BiGAT and Sophie, with and without synthetic data (*lower values are better)

Dataset	Social-BiGAT		Sophie	
	(without)	(with)	(without)	(with)
ETH	0.69/1.29	**0.65/1.19**	0.70/1.43	**0.66/1.37**
HOTEL	0.49/1.01	**0.33/0.98**	0.76/1.67	**0.63/1.42**
UNIV	0.55/1.32	**0.51/1.30**	0.54/1.24	**0.47/1.11**
ZARA1	0.30/0.62	**0.25/0.55**	0.30/0.63	**0.24/0.47**
ZARA2	0.36/0.75	0.38/0.81	0.38/0.78	**0.31/0.67**
AVG	0.48/1.00	**0.42/0.96**	0.54/1.15	**0.46/1.01**

4. Conclusion

In this paper, we have proposed a novel tool to synthesize crowd data in real buildings. Our tool leverages geometric and semantic information of BIM models, while enabling crowd simulations, using multi-agent method. The tool renders and records the crowd simulation results in the BIM model. To synthesize multiple datasets for different tasks, the corresponding information is extracted from the crowd simulation results to complete automatic annotation of the dataset. We showed that the synthetic datasets could help improve the results of crowd-related tasks and provide a research foundation for future indoor crowd-related tasks.

Our tool provides a new and efficient way to synthesize data. But it relies on artificial rules for crowd behavior, which requires the expertise of its users. In the future, we will consider data-driven methods to automatically obtain crowd action rules in different scenarios and lower the threshold for using tools.

5. Acknowledgements

This work was supported by the 2019 MIIT Industrial Internet Innovation and Development Project "BIM Software Industry Standardization and Public Service Platform".

References

Liu, S., Huang, D., and Wang, Y. (2019). Adaptive nms: Refining pedestrian detection in a crowd. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition., pp. 6459-6468.

Huang, Y., Bi, H., Li, Z., Mao, T., and Wang, Z. (2019). Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In Proceedings of the IEEE/CVF international conference on computer vision., pp. 6272-6281.

Kocabas, M., Karagoz, S., and Akbas, E. (2018). Multiposenet: Fast multi-person pose estimation using pose residual network. In Proceedings of the European conference on computer vision., pp. 417-433.

Liu, W., Salzmann, M., and Fua, P. (2019). Context-aware crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition., pp. 5099-5108.

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016, October). Playing for data: Ground truth from computer games. In European conference on computer vision., pp. 102-118.

Richter, S. R., Hayder, Z., and Koltun, V. (2017). Playing for benchmarks. In Proceedings of the IEEE International Conference on Computer Vision., pp. 2213-2222.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition., pp. 8198-8207.

Cheung, E., Wong, A., Bera, A., and Manocha, D. (2018, April). MixedPeds: pedestrian detection in unannotated videos using synthetically generated human-agents for training. In Proceedings of the AAAI Conference on Artificial Intelligence, 32.

Chai, L., Liu, Y., Liu, W., Han, G., and He, S. (2020). CrowdGAN: Identity-free Interactive Crowd Video Generation and Beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence, 01, pp.1-1.

Khadka, A., Remagnino, P., and Argyriou, V. (2020, May). Synthetic crowd and pedestrian generator for deep learning problems. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)., pp. 4052-4056.

Eastman, C., Teicholz, P., Sacks, R., Liston, K., and Handbook, B. I. M. (2008). A Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators

BuildingSMART Industry Foundation Classes (IFC). http://www.buildingsmart-tech.org/specifications/ifc-overview/. Accessed 30 Sept 2018

Sharma, S., Ogunlana, K., Scribner, D., and Grynovicki, J. (2018). Modeling human behavior during emergency evacuation using intelligent agents: A multi-agent simulation approach. Information Systems Frontiers, 20(4), pp. 741-757.

Pellegrini, S., Ess, A., and Gool, L. V. (2010, September). Improving data association by joint modeling of pedestrian trajectories and groupings. In European conference on computer vision., pp. 452-465.

Lerner, A., Chrysanthou, Y., and Lischinski, D. (2007, September). Crowds by example. In Computer graphics forum, 26, 3, pp. 655-664.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., and Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. Advances in Neural Information Processing Systems, 32.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., and Savarese, S. (2019). Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition., pp. 1349-1358.

Pellegrini, S., Ess, A., Schindler, K., and Van Gool, L. (2009, September). You'll never walk alone: Modeling social behavior for multi-target tracking. In 2009 IEEE 12th international conference on computer vision., pp. 261-268.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition., pp. 961-971.