

LiDAR-based human pose estimation with MotionBERT

Zichen Zhao¹
¹College of Artificial
Intelligence,
Hebei University of
Technology
Tianjin, China
zzcjl666@163.com

Chao Zhuang¹
¹College of Artificial
Intelligence,
Hebei University of
Technology
Tianjin, China
1494550732@qq.com

Jian Li²
²National Rehabilitation Aids
Research Center
Beijing, China
redbomb628@163.com

Hao Sun^{*1}
¹College of Artificial
Intelligence,
Hebei University of
Technology
Tianjin, China
sunhao@hebut.edu.cn

**Corresponding author*

Abstract - Human body pose estimation has a wide range of applications in the field of artificial intelligence and is gradually applied in daily life. However, for some specific scenes and people, privacy and security need to be ensured, and devices such as cameras are no longer the optimal choice, so this paper proposes the idea of using LiDAR technology to acquire human body point cloud images, and preprocessing and segmentation to extract a pure human body point cloud model. First, the useful parts around the human body are extracted by preprocessing the point cloud image obtained from LiDAR scanning. Then, a segmentation technique is utilized to separate the human body from the surrounding objects, keeping only the human body part. Next, the excess noise is removed by filtering process to get the pure human body point cloud model. Finally, the motionBert3d model is used to estimate the pose of the preprocessed point cloud image, and the pose estimation points and the 3D pose estimation model of the human body are obtained, which lays the foundation for the subsequent human motion recognition.

Index Terms - LiDAR; point cloud maps; image preprocessing; human pose estimation; motionBert3d modeling

I. INTRODUCTION

A wide variety of devices have been used for posture acquisition required for human posture estimation, such as cameras[2], physical sensors, RFID(Radio Frequency Identification), WiFi[3](Wireless Fidelity), radar[4], LiDAR (Light Detection and Ranging), etc[1].

These sensors applied to behavioral recognition are mainly classified into two categories: contact devices and contactless devices. Contact devices usually contain a variety of wearable physical sensors and wearable RFID[5] tags, etc.; compared to wireless sensors, wearable physical sensors have a natural advantage in the collection of physiological and physical information about a specific person, not only to obtain more accurate data, but also to determine the ID of a specific person directly; with the continuous development of the mobile network, the physical sensors are no longer subject to the With the development of mobile networks, physical sensors are no longer limited by distance. Non-contact devices do not require the user to wear any equipment, so there is no problem regarding the wearing of sensors. Common contactless devices applied to human posture acquisition are cameras, WiFi, RFID, radar, etc.; the use of contactless devices for human behavior recognition does not impose any additional requirements on the user, and therefore receives more attention from scholars. Different contactless devices

have different advantages and limitations. Images and videos are the most common carriers for recording human behavior, but ordinary cameras have shortcomings such as insufficient dark-light resolution, limited distance, and narrow field of view, which lead to poor robustness of the human pose estimation system based on videos or images; In addition very important point is that the camera is easy to violate personal privacy, bring security problems. RFID tags are affected by the range limitations, not suitable for a wide range of gesture tracking, can only be deployed within the scope of the tag to monitor the behavior; and RFID on non-metallic objects have limited penetration, may be affected by the human body, walls and other obstacles. WiFi signals are susceptible to interference from other wireless signals in the environment and may be used for unauthorized tracking and surveillance, raising security and privacy concerns. Radar does not need to deploy tags when identifying human behavior within a limited range, and different human behaviors have different effects on wireless signals; however, it is weakly resistant to environmental interference and susceptible to multipath effects.

LiDAR is a radar system that senses the orientation, velocity and other characteristic quantities of a target by emitting laser beams, and is currently used in the field of human behavior recognition for localization or tracking[6], and few researchers have applied it to human limb posture estimation. Lidar has a natural advantage in the field of human posture estimation due to its wide coverage area, fast data transmission, and independence from lighting conditions[7];It is possible to monitor the environment in a close range situation; it is also possible to realize the recognition of various human postures and movements in a medium and long range; Georgi[8] proposed a new method called BoxGraph to utilize 3D LiDAR data for semantic place recognition and pose estimation, which can better cope with problems such as occlusion and noise in complex scenes and is more efficient in practical applications. Cong[9] proposed an easy-to-deploy, light-insensitive approach for 3D multiplayer pose estimation in large-scale scenes that utilizes multimodal input data (images and point clouds) and temporal information to guide the network to learn natural and coherent human movements without relying on 3D pose annotations, is self-supervised using geometric constraints on the point cloud and is weakly supervised by using 2D key-points on the image. Li[4] et al. provided a deep learning-based radar human Behavior Recognition overview of radar-based human behavior

recognition techniques, radar systems and deep learning. They elaborated a deep radar activity recognition learning method based on radar echoes in one, two, and three dimensions. The experimental results demonstrate the feasibility of applying deep learning to radar-based human recognition. Luo[10] et al. used 2D LiDAR for simultaneous activity recognition of multiple people, using LiDAR clustering to identify people in the scene then tracking their trajectories using Kalman filtering and finally tracking them for a long period of time classifying the images using short term memory networks (LSTM) and temporal convolutional networks (TCN). Hasan [11] used 2D LiDAR instead of the camera to track the path of human activities by monitoring human ankles, which greatly protects the privacy and security of the user.

In this paper, 3D LiDAR is used for human posture estimation, LiDAR sensors can achieve centimeter-level positioning accuracy in point cloud localization, which is much more accurate than other wireless devices. In order to improve the LiDAR based human posture acquisition capability against environmental interference, this paper performs straight pass filtering, point cloud segmentation and outlier filtering on the collected point cloud data and then uses motionbert3d model for human posture estimation.

II. HELPFUL HINTS

In the preparation phase of the experiment, we set up an experimental platform in the school laboratory and deployed a Livox Mid-70 LiDAR sensor. On the Ubuntu operating system, we built a point cloud image acquisition system. Through this system, we successfully acquired 3D human posture point cloud image data from several participants, including fellow students and mentors, covering a wide range of movement postures including standing, sitting, walking, squatting, and lying flat and sideways. As shown in Fig 1-3.



Fig.1 Livox Mid-70 LiDAR



Fig.2 Laboratory Scene

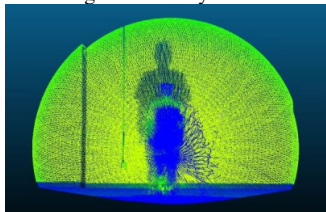


Fig.3 Lidar scanning point cloud data acquisition

Next, I will preprocess the collected point cloud image data. Preprocessing is a critical step designed to prepare the data for subsequent attitude estimation tasks.

III. EXPERIMENTATION

In normal living environments, the human body often coexists in the same space with surrounding objects such as tables, chairs, walls, floors, and other mechanical devices. When using point cloud image technology to acquire such a scene, what we get will be a series of data points in three-dimensional space, which contain not only the information of the human body, but also the information of the surrounding objects. In order to be able to accurately separate and extract the human body information from these mixed data for subsequent processing and analysis, we need to perform a series of preprocessing steps on the collected point cloud images. The preprocessing process is demonstrated using the example of a seated human posture.

A. Point cloud image preprocessing

1) Locked body area processing

Lock the human body region processing, first extract the parts around the human body in the whole big environment, PassThrough filter by setting the filtering conditions, traversing the point cloud, retaining or removing the points that meet the conditions, and finally outputting the filtered point cloud. Filter out the unwanted parts with PassThrough filter as shown in Fig.4.

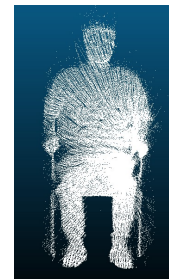


Fig.4 Point cloud plot after straight-through filtering

2) segmentation

The presence of point cloud data in the seat region is also a great challenge for human modeling. It will affect the human posture estimation and subsequent action recognition. In this paper, the region of interest (ROI) is selected for the human body point cloud data after straight pass filtering, as shown in Fig. After that, the human model region needs to be feature extracted to separate the chair from the human body.

The RANSAC[12] algorithm is based on the assumption and selection of the model, in this paper, firstly, a number of subsets in the human body point cloud are randomly selected to fit the 3D point cloud several times to estimate the parameters of this model, and through continuous iteration, the point cloud data is divided into the in-bureau 3D point cloud and the out-of-bureau 3D point cloud (outliers) by using through the set threshold value. Next, the intra-local 3D point cloud is iteratively calculated and finally the optimal geometric 3D model is obtained. The combination of RANSAC and least squares is carried out so that the advantages of both can be combined to get more ideal results.

The least squares method is the most commonly used method for solving curve fitting problems. The basic idea is to make:

$$f(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + \dots + a_m\varphi_m(x) \quad (1)$$

where φ_k is a pre-selected set of linearly independent functions, a_k is a coefficient to be determined, and the fitting criterion is to minimize the sum of the squares of the distances δ_i between $y_i (i=1,2,\dots,n)$ and $f(x_i)$, known as the least squares criterion.

Steps for using the ransac algorithm in conjunction with the least squares method:

1.First, three points in the point cloud are randomly selected and their three vectors are cross-multiplied to compute the normal vectors of the plane where the three points are located:

$$\vec{n} = (p_2 - p_1) \times (p_3 - p_1) \quad (2)$$

2.The parameters of the seat back and seat plane models were predicted by the minimum variance;

3.After obtaining the planar model parameters, the distance from a random point in the point cloud to the two planes is calculated separately:

$$d_i = \frac{\vec{n}^T (p_i - p_l)}{\|\vec{n}\|_2} \quad i = 4, 5, \dots, N \quad (3)$$

p_i is any point in the point cloud.

4.A threshold τ is pre-set so that it satisfies $d_i < \tau$ and the point cloud is filtered to meet the requirements;

5.Repeat steps 1~4, after which the point cloud with the highest number of points is saved, and an accurate plane is optimally fitted to the seating point cloud using the least squares method, and the number of iterations T is calculated as follows:

$$T = \frac{\log(1-p)}{\log(1-(1-e)^s)} \quad (4)$$

where e is the percentage of non-compliant points in the human body point cloud data in the total point cloud, s is the number of points selected in each iteration, and p is the probability that a compliant point is selected at least once.

In this paper, we calculate the distances from all point cloud data to the seat plane and find the standard deviation of these distances as a threshold for segmentation. Then, the RANSAC algorithm is applied to separate the human body point cloud from the seat plane. The result of human body information feature extraction is shown in Fig5. The advantage of this algorithm is that it can effectively separate the human body from the seat plane in different postures, and the extracted human body point cloud is of high quality, which not only retains the feature information, but also ensures the clarity of the point cloud boundaries, which is conducive to the subsequent estimation of the human body posture.

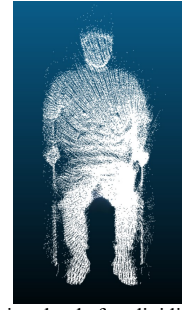


Fig.5 Point cloud after dividing the seat

3) outlier filtering processing

This hybrid filtering algorithm consists of the Statistical Outlier Removal filtering algorithm[13] and the Radius Outlier Removal filtering algorithm.

Outliers are data points that are significantly different from the neighboring points in terms of attribute values, and this difference may be due to the irregular shape of the object surface or other environmental factors. Noise points, on the other hand, are generated by random errors caused by sensors or external factors, and these points are usually scattered and few in number, located outside the main point cloud structure. The presence of outliers and noise points interferes with the accurate extraction of target features and further segmentation of the point cloud. The statistical outlier elimination filtering algorithm calculates the difference between the average distance of each point and that within its neighborhood by analyzing the distribution density of the point cloud in 3D space. Based on set criteria, this difference is evaluated and points exceeding a threshold are considered as noise or outliers and removed from the dataset, retaining only those points that meet the criteria to update the original dataset.

The steps of the statistical outlier elimination filtering algorithm are:

1.For point cloud data in space, the average distance d_{p_i} from each point to its nearest K-neighboring points is calculated by the following equation:

$$d_{p_i} = \frac{1}{K} \sum_{i=1}^K d_i \quad (5)$$

where d_i is the distance between a point in space and each point in its K-neighborhood, and K is the number of points in the neighborhood.

2.Assuming the existence of a Gaussian distribution of its mean and standard deviation, solve for the expectation d_h and standard deviation s of the mean distance between the K-neighborhoods of each point, respectively:

$$d_h = \frac{1}{n} \sum_{i=1}^n d_{p_i} \quad (6)$$

$$s = \sqrt{\frac{\sum_{i=1}^n (d_{p_i} - d_h)^2}{n}} \quad (7)$$

where n is the total number of point clouds and d_{p_i} is the average K-neighborhood distance of each point.

3. Combining the expectation d_h and standard deviation s of the average distance, the distance threshold d_l can be found:

$$d_l = d_h + \lambda \cdot s \quad (8)$$

In the formula, λ represents the relaxation parameter, the size of which is the key to the value of the threshold, and when making the parameter selection, it is necessary to take into account the characteristics of the point cloud as well as the degree of noise and other factors to ensure that the filtering effect is optimal.

4. Consider all the values of the average distance d_h with respect to the threshold d_l and compare them, after which the data larger than the threshold d_l are considered as outlier noise points to be removed.

The Radius Outlier Removal filtering algorithm counts the number of points within that radius that are neighbors of each point in the point cloud by setting a radius (neighborhood radius). If the number of neighbors of a point is less than a set threshold, the point is considered an outlier and is removed from the point cloud.

The steps of Radius Outlier Removal (ROR) filtering algorithm are:

1. set the desired radius r and the quantity threshold H ;
2. for each point P_i in the space, calculate its distance $d_{(P_i, P_j)}$ to all other points, where P_j is the other points in the point cloud.
3. Use the Kd-Tree search algorithm to search for neighboring points within a radius r for each point in the point cloud.
4. For each point P_i , count the number of neighboring points found within the radius r , denoted N_i .
5. $N_i \leq H$ is filtered out.

By removing these outliers, the quality of the point cloud data can be improved, making it more suitable for subsequent processing and analysis. The results are shown in Fig 6.

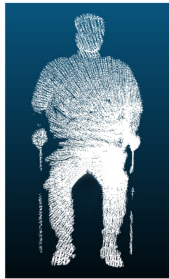


Fig.6 Point cloud map after composite outlier filtering

4) Dataset production

The LIDAR used for data collection in this experiment is Livox mid-70, which has a field of view of 70.4° (circular); the distance accuracy is 2cm;

Its field of view is 70.4° (circular), distance accuracy is 2cm, angular accuracy is less than 0.1°, data rate is 100000 points/sec, and false alarm rate is less than 0.01%. It adopts a non-repeating scanning integration method, and the length of a frame can be set to 50ms, 100ms, 200ms, 500ms, 1000ms and 3000ms, and the field of view coverage rate approaches 100% with the increase of integration time.

There is no well-known public dataset in the field of LIDAR human posture recognition, and Livox LIDAR has not been applied to the field of human posture recognition before.

Livox LIDAR has not been applied in the field of human gesture recognition before. The dataset used in this paper was collected in the laboratory of the Science Park, which is a complex environment containing conference tables, chairs, lockers, projectors and other experimental equipment, which can effectively simulate the real environment.

The point cloud data were collected during the daytime with sufficient light, and then the point cloud was converted into images to produce LiDAR-based human gesture recognition dataset. Five volunteers sequentially made five common human body postures in the LiDAR field of view, namely, standing, sitting, squatting, walking, and lying, each posture about 100 times, including almost all the positions in the room; and in the process of acquisition, the positions of the indoor conference table, seats, and lockers were changed to simulate the state of the environment when it was changed.

B. Human pose estimation using motionbert3d modeling

MotionBERT[14] is a pre-training model for video understanding and generation. It is based on the Transformer architecture, and in the pre-training phase, a motion encoder is trained to recover potential 3D motion from noisy partial 2D observations. The motion representation obtained in this way contains geometric, kinematic and physical information about the human motion. The motion encoder is implemented using a Dual-stream Spatio-temporal Transformer (DSTformer) neural network. It is able to comprehensively and adaptively capture the long-term spatio-temporal relationships between skeletal joints to obtain the 3D pose estimation with the lowest error, and the workflow is shown in Fig7.

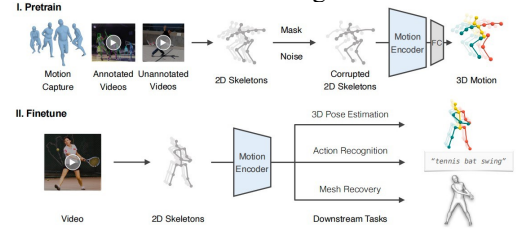


Fig.7 MotionBERT Framework overview

MotionBERT has higher efficiency and accuracy compared to traditional video understanding models. It can process both motion and appearance information in the video to better understand the video content. In addition, MotionBERT can generate high-quality videos, making it promising for a wide range of applications in video editing and generation. MotionBERT outperforms the state-of-the-art

methods in 3D pose estimation, skeleton-based motion recognition, and mesh recovery, respectively. Therefore we use MotionBERT model for human pose estimation.

The comparison of the human pose estimation 3d model for each pose generation is shown below:

1) *stand*

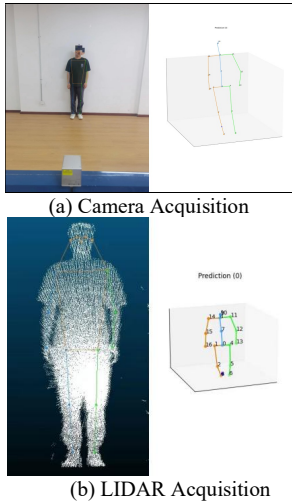


Fig.8 Comparison of Human Posture Estimation by Camera and LiDAR Acquiring Human Standing Posture Information

2) *sit*

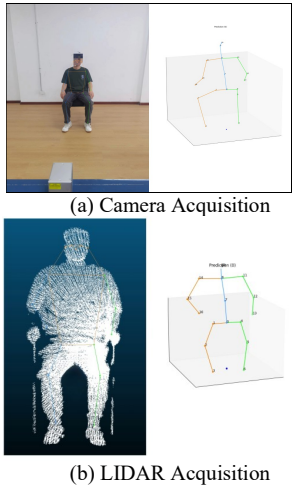


Fig.9 Comparison of Human Posture Estimation with Camera and LiDAR Acquiring Human Sitting Posture Information

3) *walk*

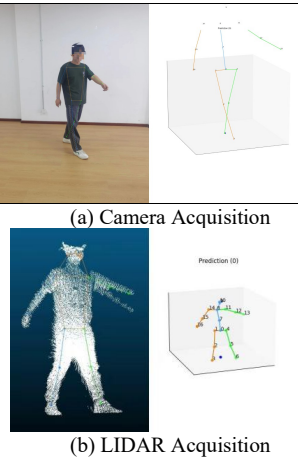


Fig.10 Comparison of Human Posture Estimation with Camera and LiDAR Acquiring Human Walking Posture Information

4) *squat*

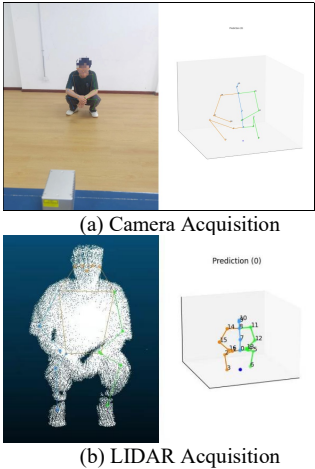


Fig.11 Comparison of Human Posture Estimation by Camera and LiDAR Acquiring Human Squatting Posture Information

5) *lie flat*

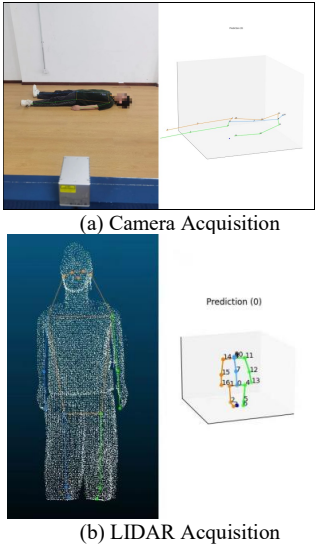


Fig.12 Comparison of Human Posture Estimation with Camera and LiDAR Acquiring Information on Human Lying Posture

6) *lie down*

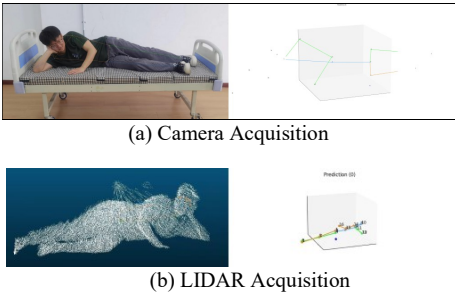


Fig.13 Comparison of human pose estimation by camera and LiDAR for obtaining human side-lying pose information

IV. EXPERIMENTAL RESULTS AND EVALUATION

Based on the experimental analysis in the previous section, we can see that: when using the MotionBERT model for human posture estimation, for the basic postures such as standing, sitting, walking, squatting, and lying down, the

model is able to accurately predict the key joint points, including the positions of the head, both shoulders, elbows, both hands, upper torso, knees, and both feet.

By comparing with the pose estimation methods based on camera captured images, we find that the MotionBERT model based on LIDAR scanning is equally capable of providing accurate pose estimation results in the aforementioned poses. In addition, since LiDAR scanning does not involve the direct capture of an individual's image, it offers a significant advantage in protecting personal privacy.

The main current problems include: 1) When the body pose is lying on the side, the model only recognizes the pose of the upper body, and fails to generate complete pose estimation points for both arms. This suggests that the model may have limitations in dealing with specific poses, especially when dealing with partially occluded body or complex poses. 2) The filtered point cloud data of the side-lying body does not achieve the expected ideal effect, and some noise is still left. This may be due to the parameter settings of the filtering algorithm or the algorithm's own insufficient processing ability for specific types of noise, resulting in the failure to completely remove the noise, which affects the accuracy of the pose estimation.

V. CONCLUSION

In this paper, we utilize LiDAR technology to scan the human body and acquire point cloud data, followed by a series of preprocessing steps, including the use of straight-through filtering to extract the useful parts around the human body and the use of the RANSAC algorithm to efficiently segment the human body from the surrounding objects, and in addition, in order to further improve the quality of the data, we have also performed outlier filtering of the extracted human body point cloud, so as to obtain a pure human body point cloud model, and the dataset was produced in this way. After completing these preprocessing steps, we applied the MotionBERT 3D model to the processed point cloud images for pose estimation, and successfully obtained accurate pose estimation points and 3D pose estimation models of the human body for the basic poses such as standing, sitting, walking, crouching and lying down.

The innovation of this experiment lies in the fact that LIDAR technology has significant advantages in protecting personal privacy compared to other devices for acquiring human body postures. Since LiDAR scanning does not involve the direct capture of an individual's image, there is no leakage of personal privacy information during the scanning process. This feature allows the MotionBERT 3D human pose estimation method based on LiDAR point cloud images to ensure the accuracy of the estimation while effectively protecting personal privacy.

Future work: 1) Enhance the preprocessing to remove more noise and obtain a purer human body point cloud model 2) Make changes to the motionbert3d model to include spatial location information in the channel attention, such as the Coordinate Attention (CA) mechanism, which can better capture spatial correlations in the image 3) Utilize the

produced dataset to perform model training, adding more samples of side lying or similar extreme poses to the training data to allow the model to learn more diverse variations of human postures.

ACKNOWLEDGMENT

This work was supported by National Key Research and Development Program (No. 2020YFC2007700).

REFERENCES

- [1] Y. Zhu, J. Zhao, Y. Wang, and B. Zheng, "A review of human behavior recognition algorithms based on deep learning", *Acta Automatica Sinica*, vol. 42, no. 6, pp. 848-857, June 2016.
- [2] L. Qin, Q. Hu, Q. Huang and Q. Tian, "Motion recognition based on feature point locus", *Chinese Journal of Computers*, vol. 37, no. 6, pp. 1281-1288, 2014.
- [3] J. Huang, B. Liu, P. Liu, C. Chen, N. Xiao, C. Zhang, et al, "Towards Anti-interference WiFi-based Activity Recognition System Using Interference-Independent Phase Component", in *IEEE International Conference on Computer Communications*, Jul. 2020, pp. 576-585.
- [4] X. Li, Y. He and X. Jing, "A Survey of Deep Learning-Based Human Activity Recognition in Radar", *Remote Sensing*, vol. 11, no. 9, pp. 111-119, April 2019.
- [5] J. Wang, O. Abari and S. Keshav, "Challenge: RFID Hacking for Fun and Profit", in *Annual International Conference on Mobile Computing and Networking*, Oct. 2018, pp. 461-470.
- [6] K. Koide, J. Miura and E. Menegatti, "A portable three-dimensional LIDAR-based system for long-term and wide-area people behavior measurement", *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, pp. 1-16, 2019.
- [7] B. Galai and C. Benedek, "Feature selection for Lidar-based gait recognition", in *International workshop on computational intelligence for multimedia understanding (IWCIM)*, Oct. 2015, pp. 1-5.
- [8] G. Pramatarov, D. De Martini, M. Gadd and P. Newman, "BoxGraph: Semantic place recognition and pose estimation from 3D LiDAR", in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 7004-7011.
- [9] P. Cong, Y. Xu, Y. Ren, J. Zhang, L. Xu, J. Wang, et al, "Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Jun. 2023, pp. 461-469.
- [10] F. Luo, S. Poslad and E. Bodanese, "Temporal Convolutional Networks for Multiperson Activity Recognition Using a 2-D LIDAR", in *IEEE INTERNET THINGS*, Mar. 2020, pp. 7432-7442.
- [11] M. Hasan, J. Hanawa, R. Goto, H. Fukuda, Y. Kuno and Y. Kobayashi, "Tracking people using ankle-level 2D LiDAR for gait analysis", in *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2020 Virtual Conferences on Software and Systems Engineering, and Artificial Intelligence and Social Computing*, Springer International Publishing, Jul. 2020, pp. 40-46.
- [12] R. Raguram, O. Chum, M. Pollefeys, J. Matas and J. Frahm, "USAC: A universal framework for random sample consensus", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no.8, pp. 2022-2038, 2022.
- [13] X. Liu and Y. Song, "Multi-scale point cloud noise elimination by filtering method", *Beijing Surveying and Mapping*, vol.36, no.11, pp. 1486-1489, 2022.
- [14] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu and Y. Wang, "Motionbert: A unified perspective on learning human motion representations", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15085-15099.