

An Accurate Point Cloud-Based Human Identification Using Micro-Size LiDAR

Shota Yamada

Grad. Sch. of Info. Sci. and Tech.
Osaka University
Osaka, Japan
sho-yamada@ist.osaka-u.ac.jp

Hamada Rizk

Grad. Sch. of Info. Sci. and Tech.
Osaka University Osaka, Japan
Tanta University Tanta, Egypt
hamada_rizk@f-eng.tanta.edu.eg

Hirozumi Yamaguchi

Grad. Sch. of Info. Sci. and Tech.
Osaka University
Osaka, Japan
h-yamagu@ist.osaka-u.ac.jp

Abstract—The demand for safety-boosting systems is increasing, especially nowadays, to limit the rapid spread of COVID-19. Real-time life-logging is an essential application towards tracking infected cases and thus containing the pandemic outbreak. This application raises the need for an accurate human identification technology where cameras cannot be adopted due to privacy. Recently, LiDAR sensor has attracted attention due to its ability to represent the surrounding environment in the form of 3D point cloud. In this paper, we introduce a novel wearable device of a micro-size LiDAR to build an onboard human identification system for life-logging. The system acquires 3D point cloud data of the surrounding environment from which subject-discriminative signatures are extracted. This is achieved by removing noise and background using Spatio-temporal density clustering and fisher vector representations. The extracted features are then used to train a random forest classifier for subject identification. We have implemented and evaluated the proposed system on six different subjects. The results show that the proposed system can effectively remove noise and background and accurately identify subjects with 99.9% accuracy.

Index Terms—point cloud-based recognition, LiDAR, Human identification, user tracing for COVID-19

I. INTRODUCTION

In an attempt to contain the rapid spread of COVID-19 by tracking infected cases and ensuring social distance preservation, academia and industry seek to develop new technologies that combine human identification techniques and geolocation features [1]. These technologies are also aimed to provide helpful information that facilitates understanding the spread of the virus along with the individual practices to avoid the infection spread via life-logging tools. Moreover, this technology is highly required in public closed environments such as kinder gardens, schools, and universities.

Although the prominence of proximity-based solutions, e.g., Bluetooth, they highly rely on the availability of Bluetooth-enabled smartphones, which are not always available specifically in education institutes. As an alternative, a lot of work has been devoted based on state-of-the-art technologies in human identification using biometric information such as the face [2] and voice [3]. For this, computer vision techniques have been leveraged for recognizing faces on based building infrastructure cameras. This line of research limits the feasibility of aimed system due to the limited coverage of fixed cameras. Additionally, voice-based human identification leveraging the

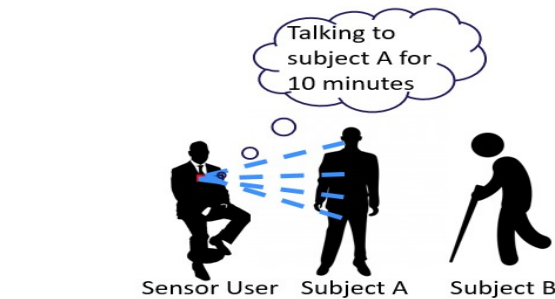


Fig. 1. The typical situation of our system usage

onboard microphone at the user's smartphone. However, this involves recording the user's speech for identification which may include sensitive/private information. Unsurprisingly, various privacy issues have arisen due to using cameras or microphones in all environments, e.g., restrooms.

Light Detection and Ranging can detect surrounding objects using eye-safe lasers (class 1A, near-infrared spectrum) and determine distances to target objects with sub-decimeter errors. It is usually used in large indoor environments such as malls, museums, and governmental buildings security surveillance, people counting, tracking in public space, path planning, marketing, etc. LiDARs capture 3D point cloud data of 3D dimensional coordinates points that do not contain personal information.

In this paper, we propose a system for enabling human identification with a cheap wearable LiDAR, as shown in Fig. 1. We introduce the first micro-size LiDAR device that can scan and represents the surrounding environment in a privacy-preserving point cloud representation. This representation can be leveraged to identify the user contacting the sensor wearer without revealing any sensitive information such as visible images or voices. Specifically, the proposed system seamlessly constructs a signature database for the considered subjects to identify. These signatures are extracted from the captured point cloud representations of each user using a computationally efficient fisher vector-based representation. Then a random forest classifier is trained to identify the user in question.

Nevertheless, the proposed system needs to make provision for a number of challenges. Firstly, extracting meaningful context is not a straightforward process due to its unordered,

unstructured, and varying size nature, unlike visible images. Secondly, the limited sensing capability of the proposed micro-size LiDAR in terms of range and noise sensitivity. Finally, considering the limited computational power at the edge device. To tackle these issues, we introduce a number of innovative modules, including noise detection and removal using spatial-temporal density-based clustering over cascaded point cloud frames. Fixed-size features are extracted using the fisher vector method leveraging its implicit symmetric functions. The extracted features are then leveraged to train a computationally efficient classifier for human identification purposes.

The proposed system has been evaluated on six different subjects of students and staff in our lab. The results show that the proposed system provides a human identification accuracy of 99.9%. We also show the efficacy of noise removal when multiple consecutive frames are considered. These results highlight the proposed system as the next-generation human identification technology.

The rest of this paper is structured as follows. We carry out a literature review in Section II. In Section III, we provide an overview of the proposed system, while Section IV presents its details. In Section V, we describe the data collection process and how the system is tested. Finally, we conclude the paper in Section VI.

II. RELATED WORK

In this section, we discuss the most relevant literature to the proposed system.

A. Human Identification

Many researches have been proposed for human-centric applications e.g., tracking [4]–[14] or identification [2], [3], [15]. Due to the pandemic, human Identification attracted a lot of attention relying on different concepts. For instance, humans can be identified based on face recognition [2], gait recognition [15], and voice recognition [3]. Face and voice recognition have shown superior performance for discriminating different subjects. However, these techniques usually yield privacy concerns that hinder their ubiquitous adoption in practice.

The face recognition systems have initiated with the Eigen-face [16] approach, which obtains a lower-dimensional representation through certain distribution assumptions. However, this approach fails to deal with the uncontrolled facial changes that stray from their prior assumptions. Learning-based local descriptors were proposed, [17], [18] which could not wholly ensure robustness against facial changes. Deep learning techniques have emerged and become common after AlexNet [19] which achieved the best accuracy in the ImageNet competition at that time. In 2014, DeepFace [20] achieved 97.35% accuracy on the LFW benchmark, which is very close to human performance (97.53%). Then, many researchers focused on deep learning, which has achieved state-of-the-art performance so far with accuracy up to 99.80%.

Gait-based human recognition gets a lot of attraction from many fields such as computer vision. This is done by extracting features from the subject's body during movement. For this,

the system in [21] uses 3D convolutional neural networks to learn the gait from different viewing angles. While a person's skeletons and joint features are learned using a time-based graph LSTM network [22]. However, this line of research involves many challenges, including the variability of gait patterns and the recognition difficulty in crowded environments.

For human(speaker) recognition by voice, many techniques [23], [24] have been proposed leveraging the powerful capability of deep learning. For instance, the system in [23] proposed the concept of d-vector to boost the speaker's recognition. This is done by training a model to extract features from which the d-vector is computed, enabling speaker identification. In [24], an encoder-discriminator architecture method was proposed for learning speaker embeddings. This method shows that the model can learn speaker representations efficiently and perform better than the baseline techniques (i-vector and t-vector).

On the contrary, the proposed system relies on a 3D point cloud that maintains users' privacy compared to vision-based or acoustic techniques. This suits many use cases where privacy is at a premium, e.g., fitting rooms in shops.

B. Point cloud-based learning

Recent years have witnessed the prominence of LiDAR-based (i.e. 3D point cloud-based) applications in many fields such as autonomous driving, archaeology, agriculture, etc. In autonomous driving, LiDAR is used for object detection like pedestrians and other vehicles. A high-quality LiDAR of wide-range and 360 degrees angles is attached on top of cars to scan the surrounding information enabling a better understanding of the environment. There are two techniques for object detection. i.e. Region proposal-based methods and single-shot methods. Region-proposal-based methods create several regions which contain objects and then extract features for each region to determine which class the objects belong to. There are three ways for these methods: *Multiview-based*, *Segmentation-based* and *Frustum-based methods*. MultiView-based methods [25], [26] fuse features for each proposal from different view maps such as 2D image and bird's eye view (BEV). Segmentation-based methods [27]–[29] use semantic segmentation techniques by dividing points into foreground points and background points. Then, background points are removed and high-quality proposals are generated only for foreground points for saving the computational cost.

Frustum-based methods [30] use 2D object detectors for proposing 2D candidate regions of objects and generate a 3D frustum proposal for each 2D candidate region. It works efficiently for proposals of possible locations of objects. Single-shot methods use a single-stage network and directly predict probabilities of each class and generate 3D bounding boxes of objects. Their processing speed is higher than region proposal-based methods because they do not have a stage for region proposal. Single-shot methods can be divided into three categories based on the type of input data: BEV-based, discretization-based and point-based methods. BEV-based methods [31] [32] take the bird's eye view (BEV) representation as inputs.

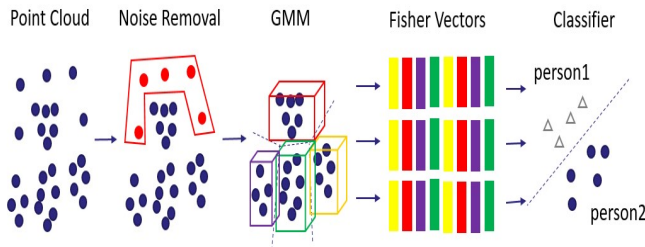


Fig. 2. System flow over the different modules.

Discretization-based methods [33] [34] convert point clouds into a certain fixed discrete representation and predict object categories and 3D bounding boxes by applying Convolutional Neural Network(CNN). Point-based methods [35] take raw point clouds as inputs.

Deep learning-based methods can be categorized into indirect and direct ways [36]. In indirect ways, point clouds are converted into a regular structure such as multi-views [37] and voxel grids [38]. While direct ways exploit raw point clouds e.g PointNet [39] which directly takes raw point cloud as inputs for classification and segmentation. In this paper, they address the common challenges of point clouds. i.e., invariance of permutation and transformation of the point cloud. All these methods are computationally intensive and require high computational resources, hindering their adoption at edge devices.

Different from these techniques, this is the first work to consider the human identification-based 3D point cloud of LiDAR device. Additionally, the proposed system is designed to be computationally efficient facilitating its adoption on wearable LiRAR as an edge device.

III. SYSTEM OVERVIEW

The proposed system consists of two main stages: **the sensing stage** and **the learning stage**. The sensing stage involves the design of a brand new micro LiDAR, which we call Hitonavi- μ . The sensor is designed to be worn by the end-user as a necklace (around her neck) and outputs 3D point clouds for the surrounding environment. These point cloud scans initiate the learning stage for further processing. In practice, we collect point cloud scans for different subjects while standing. The learning stage aims at extracting salient features (signature) for each considered subject facilitating the subject identification process. Therefore, this stage begins by collecting point cloud data for each subject from which a signature database will be extracted. Then, we build a noise and background removal module using density-based clustering over consecutive scans. This is done by integrating multiple consecutive scans to increase the density of the points, enabling better recognition of shapes, objects, and noise. Next, discriminative features are extracted from the noise-free point cloud scans using a feature extraction module that leverages Fisher vector representation. The output of this stage is the signature database for all subjects. Finally, a machine learning classifier is trained to identify surrounding users in the

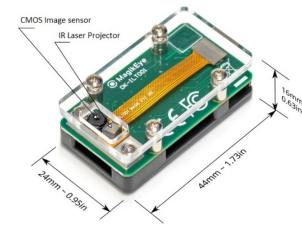


Fig. 3. The micro size LiDAR used at the core of the proposed system.



Fig. 4. Sensor and compute module.

environments given fisher vector representation of the scanned point cloud.

IV. THE SYSTEM DETAILS

This section presents the details of the different modules of the proposed system including the sensing and data collection stage and the learning stage. Fig. 2 shows the system architecture, including its different modules.

A. The Sensor stage

This section describes the characteristics of Hitonavi- μ (the proposed sensor). Hitonavi- μ a battery-equipped micro version of Light Detection and Ranging (LiDAR) together with its computing unit. LiDAR, in general, is remote sensing means used to measure the distance to the surface of surrounding objects. It measures the time for a laser pulse to travel from the sensor to the surrounding objects/surfaces and return back and thus estimates the distance. Therefore, it is intensively used in many applications, including the autonomous cars industry [36] to detect the presence of obstacles and indoor navigation for tracking humans and robots [40]. All these applications are built on large-size LiDAR with powerful sensing capabilities. Thus, it comes with a high cost, which hinders their widespread adoption as a privacy-preserving technology. Therefore, in this paper, we introduce a micro version of the LiDAR sensor¹ (shown in Fig. 3) which is relatively smaller and lighter compared to the consumers' LiDAR devices. It has Width, depth, and height of is 44, 24, and 16, respectively. The rationale behind this design is to facilitate leveraging the sensor as a wearable device, paving the road for the next generation of privacy-preserving vision technology. The LiDAR sensor is connected to a computing unit of battery-equipped Raspberry Pi 4 Model B [41] for processing our algorithms. Raspberry Pi 4 Model B is a compact single board

¹This sensor was manufactured by MagikEye company.

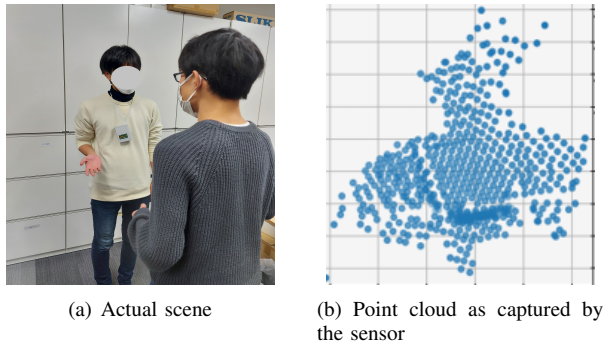


Fig. 5. Typical use case scenario.

computer that can be used for mobile devices having a CPU of ARM Cortex-A72 quad-core at 1.5GHz and 4GB RAM. Hitonavi- μ composed the micro-size sensor and the compute module is shown in Fig. 4. The average frame rate of the sensor equals 30 fps.

B. The Data Collection

Here, we collect the 3D point cloud of different subjects to construct a signature database. This is done by capturing the scene using the sensor and uploading the acquired point cloud to our server through the onboard WiFi module of Hitonavi- μ . The data is captured while subjects are in conversation (stationary), enabling the aimed life-logging application. The sensor is worn around the neck on the subject's chest, capturing the surrounding objects and subjects at a distance of up to 3m. Fig. 5 shows the data collection environment setup. Fig. 5(a) shows a typical use case of the sensor, while Fig. 5(b) depicts its corresponding captured point cloud. As can be seen from the figure, the point cloud data has a well representation of the subject with a 3D point cloud. Each point of the cloud has only x, y, and z coordinate information. It can also be seen the distortion around the head due to the presence of some noisy points.

C. The Learning Part

The learning stage consists of three modules: Noise Removal, Feature Extraction, and Subject Identification.

1) *Noise Removal*: This module aims to remove the undesirable noise and background that may negatively affect the quality of the recognition as it deceives the model. For detection and removal of noise, we leverage the Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) [42]. DBSCAN has a number of advantages that motivated us to leverage it for the purpose of noise removal. Firstly, it can work independently of the domain knowledge information e.g. the number of clusters. This makes it a better choice as compared to traditional clustering techniques such as K-Means [43]. Additionally, unlike other clustering techniques, DBSCAN can create clusters of arbitrary shapes and of varying densities. Finally, DBSCAN is robust to outliers and noise and can efficiently detect them [44], [45].

The main idea of DBSCAN is that it works on the assumption that clusters are dense regions in space, such as

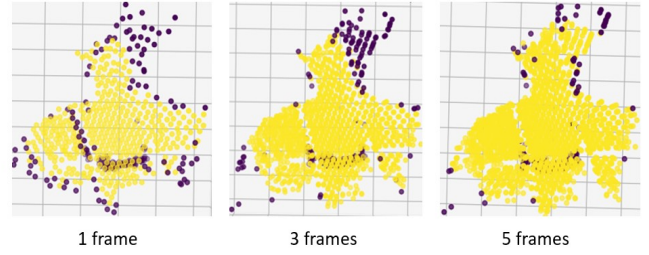


Fig. 6. The effect of the Noise Removal module on the input point cloud frame. The Noise removal is applied on 1,3 and 5 integrated frames.

the subject's arm, chest and head, separated by regions of lower density. Therefore, DBSCAN can identify clusters by looking at the local density of the points. DBSCAN requires two parameters: *Eps* and *MinPts*. *Eps* is the radius of the circle to be created around each point to check the density and *MinPts* is the minimum number of points required inside that circle for that point to be classified as a *Core* point. DBSCAN creates a circle of *Eps* radius around every point and classifies them into *Core* point, *Border* point, and *Noise*. A point is a *Core* point if the circle around it contains at least *MinPts* points. If the number of points is less than *MinPts*, then it is classified as *Border* point, and if there are no other points around any point within *Eps* radius, then it is treated as *Noise*.

DBSCAN efficacy is boosted by considering the temporal effect across consecutive point cloud frames. To do that, the points in n consecutive frames are combined, leading to an integrated frame. Thus, DBSCAN is applied to cluster points in that integrated frame. The intuition behind this process is that the integration increases the density difference between human and noise clusters facilitating noise detection and removal.

The output of this module is a number of clusters representing points of the different parts of the body (we merge them in one class) along with a number of points annotated as *Noise*. These Noise points are removed and the remaining points are fed to the next module for further processing.

2) *Feature Extraction*: In this section, we propose the feature extraction technique based on Fisher Vector (FV) representations. The FV representation's independence of sample size makes it a natural choice for representing point cloud data. Specifically, FV representation is employed to define discriminative signatures for different subjects' data of variable sizes (as in 3D point cloud frames). Additionally, it implicitly represents the spatial locality of points. This signature is defined as the deviation of 3D points from a generative model (e.g., Gaussian Mixture Model (GMM)). This can be done by calculating the gradients of the sample's log-likelihood w.r.t. the model parameters (i.e., weight, mean, and covariance). Additionally, FV maintains a fixed-size grid structure, making it a convenient input to any classifier.

3) *Subject Identification*: In this section, we describe how we build a classification model that can precisely differentiate between different subjects given the extracted Fisher representations. For this, Random Forest Classifier is employed here.

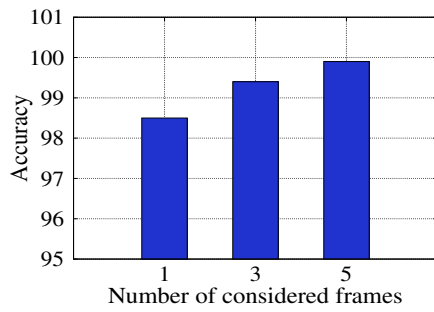


Fig. 7. The accuracy of the proposed system.

The random forest contains many decision trees; each of them outputs a class prediction, and the class with the most votes will be the considered model.

The motivation for using a Random Forest classifier is its ability to train a large number of relatively uncorrelated models (trees) operating as an ensemble of more accurate predictors than individual ones. Additionally, Random Forest is a flexible, easy-to-use supervised learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time [46]. It is also one of the most used algorithms because of its simplicity and diversity. The input to the classifier is the features obtained from the feature extraction module. The best performance has been obtained with 200 trees and the *entropy* as a split metric. Once the model has been trained, it can recognize the object in question given the extracted Fisher representation.

V. EVALUATION

We conducted an experiment to evaluate the end-to-end performance of the proposed system. The experiment involves *six* subjects, including students and staff members in our lab. The distance between the worn sensor and the target subject is arbitrary and varies based on the subjects' preferences during a natural conversation. For each targeted subject, the point cloud signatures are captured over the course of *three* minutes. This data is further divided into 80% training and 20% testing.

Fig. 6 shows the performance of the noise removal module in three cases: single frame, integrating 3 frames and integrating 5 frames. The figure confirms that the more frames to integrate the better noise detection is achieved. This can be justified as the integration increases the density of the main clusters of the subject while keeping the noise sparse at low density. This boosts the noise detection and clustering skills of DBSCAN.

Fig. 7 shows the classification accuracy of the *six* subjects when changing the number of integrated frames. As can be seen, integrating more frames enhances the overall subject identification accuracy up to 99.9% at the case of integrating *five* consecutive frames (1 second). Beyond, five frames the accuracy drops as the captured frame may include the target subject at multiple different positions in the scene leading to a different shape which is highly probable to deceive the classification model.

Finally, we evaluate per subject identification accuracy at each case of integrated frames. Fig. 8 shows the confusion

matrices of the *six* subjects when tested at each case of integrated frames. The confusion matrices show the gain earned by integrating more frames and the superiority of five-frames-based clustering. The results also confirm the validity of the proposed system for identifying different subjects even when a single frame-based clustering. This highlights the promise of the proposed system as the next generation of the privacy-preserving life-log system.

VI. CONCLUSION

We proposed a point cloud-based subject identification system for the purpose of life-logging using the first micro-size wearable LiDAR. We presented the details of the system and its ability to detect and remove noise and extract discriminative features facilitating the accurate identification of different subjects. To achieve that, the proposed system leverages a combination of spatio-temporal clustering for noise removal and a learned representation using Fisher vector. This representation is further harnessed to train a random forest classifier for recognizing the user in the scene. Implementation of our system and evaluating it with six different subjects showed that the proposed system could obtain a subject identification accuracy of 99.9%.

In the future, we plan to investigate the system performance when a larger number of subjects are considered in different human interaction situations.

ACKNOWLEDGMENT

This work was supported by JST, A-STEP Grant Number JPMJTR20RV, Japan.

REFERENCES

- [1] R. A. Kleinman and C. Merkel, "Digital contact tracing for covid-19," *CMAJ*, vol. 192, no. 24, pp. E653–E656, 2020.
- [2] G. M. Zafaruddin and H. Fadewar, "Face recognition using eigenfaces," in *Computing, communication and signal processing*. Springer, 2019, pp. 855–864.
- [3] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in labview [j].," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.
- [4] H. Rizk, H. Yamaguchi, T. Higashino, and M. Youssef, "A ubiquitous and accurate floor estimation system using deep representational learning," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 540–549.
- [5] H. Rizk, M. Abbas, and M. Youssef, "Device-independent cellular-based indoor location tracking using deep learning," *Pervasive and Mobile Computing*, p. 101420, 2021.
- [6] K. Alkief, A. Othman, H. Rizk, and M. Youssef, "Deep learning-based floor prediction using cell network information," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 663–664.
- [7] I. Fahmy, S. Ayman, H. Rizk, and M. Youssef, "Monofi: Efficient indoor localization based on single radio source and minimal fingerprinting," in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 2021, pp. 674–675.
- [8] H. Rizk, "Solocell: Efficient indoor localization based on limited cell network information and minimal fingerprinting," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 604–605.
- [9] H. Rizk, A. Shokry, and M. Youssef, "Effectiveness of data augmentation in cellular-based localization using deep learning," in *Proceedings of the International Conference on Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019.
- [10] H. Rizk, "Device-invariant cellular-based indoor localization system using deep learning," in *The ACM MobiSys 2019 on Rising Stars Forum*, ser. RisingStarsForum'19. ACM, 2019, pp. 19–23.

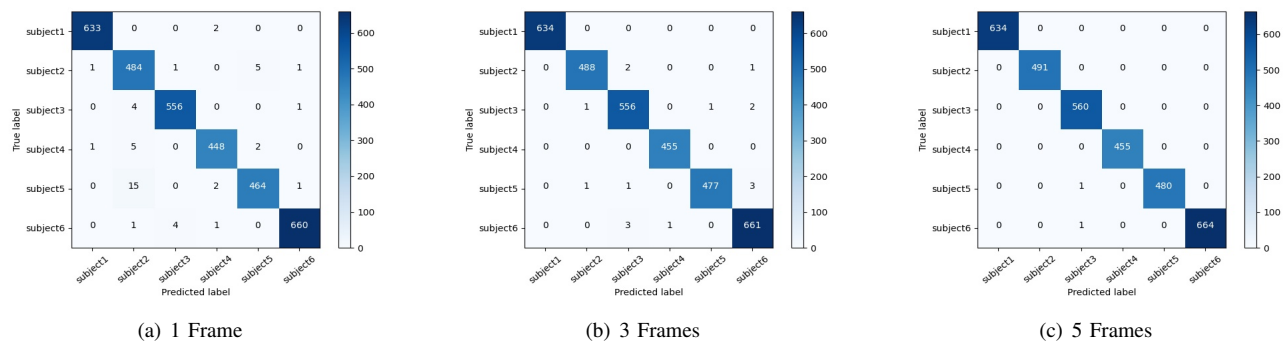


Fig. 8. Accuracy of the proposed system.

- [11] H. Rizk and M. Youssef, "Monodcell: A ubiquitous and low-overhead deep learning-based indoor localization with limited cellular information," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2019, pp. 109–118.
- [12] H. Rizk, T. Amano, H. Yamaguchi, and M. Youssef, "Smartwatch-based face-touch prediction using deep representational learning," in *Proceedings of the 18th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, EAI. Springer, 2021.
- [13] H. Rizk, M. Abbas, and M. Youssef, "OmniCells: cross-device cellular-based indoor location tracking using deep neural networks," in *the International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2020.
- [14] H. Rizk, M. Torki, and M. Youssef, "CellinDeep: Robust and Accurate Cellular-based Indoor Localization via Deep Learning," *IEEE Sensors Journal*, 2018.
- [15] J.-H. Yoo, D. Hwang, K.-Y. Moon, and M. S. Nixon, "Automated human recognition by gait using neural network," in *2008 First Workshops on Image Processing Theory, Tools and Applications*, 2008, pp. 1–6.
- [16] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [17] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2707–2714.
- [18] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 289–302, 2013.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.
- [21] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3d convolutional neural networks," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 4165–4169.
- [22] F. Battistone and A. Petrosino, "Tglstm: A time based graph deep learning approach to gait recognition," *Pattern Recognition Letters*, vol. 126, pp. 132–138, 2019, robustness, Security and Regulation Aspects in Current Biometric Systems. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865518301703>
- [23] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *the international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.
- [24] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," *arXiv preprint arXiv:1812.00271*, 2018.
- [25] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [26] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [27] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," *arXiv preprint arXiv:1812.05276*, 2018.
- [28] S. Shi, X. Wang, and H. Li, "Pointcnn: 3d object proposal generation and detection from point cloud," in *the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.
- [29] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [30] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.
- [31] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.
- [32] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning*. PMLR, 2018, pp. 146–155.
- [33] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.
- [34] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1513–1518.
- [35] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 040–11 048.
- [36] J. Zhang, X. Zhao, Z. Chen, and Z. Lu, "A review of deep learning-based semantic segmentation for point cloud," *IEEE Access*, vol. 7, pp. 179 118–179 133, 2019.
- [37] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *the IEEE international conference on computer vision*, 2015, pp. 945–953.
- [38] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *the International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 922–928.
- [39] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
- [40] H. Rizk, H. Yamaguchi, M. Youssef, and T. Higashino, "Gain without pain: Enabling fingerprinting-based indoor localization using tracking scanners," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 550–559.
- [41] "Raspberry pi foundation," <https://www.raspberrypi.com/products/raspberry-pi-4-model-b/>, accessed: 2021/11/30.
- [42] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [43] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
- [44] H. Rizk, S. Elgokhy, and A. Sarhan, "A hybrid outlier detection algorithm based on partitioning clustering and density measures," in *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*. IEEE, 2015, pp. 175–181.
- [45] A. Elmogy, H. Rizk, and A. M. Sarhan, "Ofcod: On the fly clustering based outlier detection framework," *Data*, vol. 6, no. 1, p. 1, 2021.
- [46] V. Erdélyi, H. Rizk, H. Yamaguchi, and T. Higashino, "Learn to see: A microwave-based object recognition system using learning techniques," in *Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking*, 2021, pp. 145–150.