IMPROVING 3D PEDESTRIAN DETECTION FOR WEARABLE SENSOR DATA WITH 2D HUMAN POSE

Vinu Kamalasanan[1], Yu Feng[1], Monika Sester [1*]

[1] Institute for Cartography and Geoinformatics, Leibniz Universität Hannover, 30167 Hannover, Germany

Commission IV, WG IV/9

KEY WORDS: 3D pedestrian detection, human pose estimation, augmented reality, shared space, wearable sensor

ABSTRACT:

Collisions and safety are important concepts when dealing with urban designs like shared spaces. As pedestrians (especially the elderly and disabled people) are more vulnerable to accidents, realising an intelligent mobility aid to avoid collisions is a direction of research that could improve safety using a wearable device. Also, with the improvements in technologies for visualisation and their capabilities to render 3D virtual content, AR devices could be used to realise virtual infrastructure and virtual traffic systems. Such devices (e.g., Hololens) scan the environment using stereo and ToF (Time-of-Flight) sensors, which in principle can be used to detect surrounding objects, including dynamic agents such as pedestrians. This can be used as basis to predict collisions. To envision an AR device as a safety aid and demonstrate its 3D object detection capability (in particular: pedestrian detection), we propose an improvement to the 3D object detection framework Frustum Pointnet with human pose and apply it on the data from an AR device. Using the data from such a device in an indoor setting, we conducted a comparative study to investigate how high level 2D human pose features in our approach could help to improve the detection performance of orientated 3D pedestrian instances over Frustum Pointnet.

1. INTRODUCTION

Pedestrian friendly urban designs like walkable cities and shared spaces have recently gained a lot of attention. While the former has focused more on pedestrian needs in urban and suburban environments necessitating a pedestrian network as a criterion for its successful design, the latter has emphasized more on promoting walking by mixing different traffic modes (cars, cyclists, and pedestrians) (Hamilton-Baillie, 2008) with no or reduced infrastructure. In either of these spaces, collisions are a potential safety threat considering pedestrian-pedestrian interactions (conflicts via congestion in pedestrian network (Wang et al., 2016)) or when interactions with different road users (e.g., pedestrian-car collision).

The basic idea of shared spaces is to mix traffic participants to create unclear situations to promote lower vehicle speed promoting walkability; however, the elderly and disabled feel less safe as they are expected to be more cautious. The inability to anticipate an upcoming danger due to reduced cognition or a confusion over priority while crossing paths could result in collisions that is life-threatening. Therefore, possible wearable conflict detection systems (e.g., intelligent mobility aids) need to be explored to enhance safety for these vulnerable road users (VRU). Conflict in this scope of work is inspired from (Javid and Seneviratne, 1991). It is defined as a traffic event involving one or more pedestrians and one or more vehicles, where both perform actions, such as applying changes in direction or speed, to avoid a collision.

Augmented Reality (AR) devices use perception sensors for spatial mapping to place virtual 3D content aligned with the real world space. They use RGBD sensors and can acquire both RGB images and depth information in raw format. Currently such sensors are also used for 3D pedestrian detection, a fundamental component for follow-up motion prediction and collision detection in autonomous driving. Using the sensor capabilities of an AR device, if we can realise a collision detection system where nearby pedestrian are detected and their future motion are predicted; these head mounted headsets can serve as safety aids and further the research along using 3D augmentation. This can be used for realising virtual traffic lights to follow rules (Kamalasanan and Sester, 2020) and hence traffic behaviour in shared spaces.

However, current research in 3D pedestrian detection is mainly applied to autonomous driving and robotics and there are fewer studies of data from wearable devices that serve pedestrians safety. As a first step in the direction of realising a wearable safety detection system using an AR device, we did experiments in an indoor environment with pedestrian motion, and collected data from these wearable sensors.

In our work we emphasize that using extra orientation information could improve 3D detection and propose to refine the 3D orientation of people by including 2D human pose. Using the device sensors and improved feature representation via our proposed Pedestrian Pose enhanced Frustum PointNets architecture (PPEF-PointNets), we realise a 3D pedestrian detection system. Finally, we perform extensive experiments and show that our approach achieves better results than F-PointNets for pedestrian detection.

2. RELATED WORK

This section briefly introduces the basic idea of mobility aids and reviews recent work on 3D object detection using RGBD data. In addition, the literature related to the fusion of human pose information is summarized.

* Corresponding author

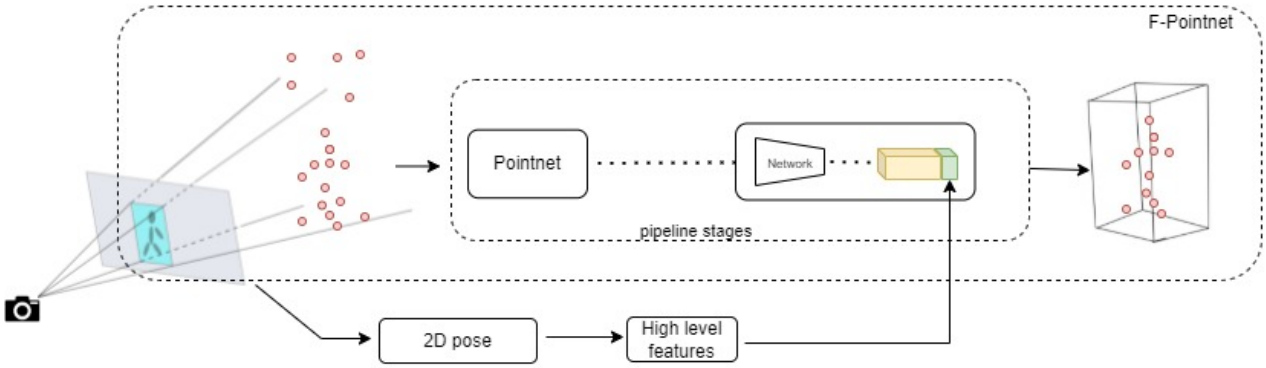

Figure 1. The whole proposed pedestrian pose based fusion approach incorporating the 2D pose features into the F-PointNets.

2.1 Intelligent mobility aids

Intelligent mobility aids for the elderly represent a class of assistive devices that attempt to augment the user's current abilities instead of replacing them and are supported by an array of sensors for environmental perception. While a larger proportion of research has focused on the white cane and wheelchair, recent works have focused along wearable systems capable of dynamically detecting obstacles (Poggi et al., 2015), with emphasis on real-time performance along with environmental perception features (e.g., crosswalk detection). Inspired from social robots, some other works have taken into account the context (Ito and Kamata, 2013) while predicting conflicts in pedestrians environment but are limited to a wheelchair approach and lack the 3D perception system. Many of these systems use auditory or haptic feedback to warn collisions and lack any visual interfaces.

2.2 3D pedestrian detection using RGBD Data

Based on RGBD data, (Kollmitz et al., 2019) extended a Faster R-CNN model (Ren et al., 2015) to regress the 3D centroids of pedestrians. Meanwhile, (Linder et al., 2020) extended the YOLO v3 model (Farhadi and Redmon, 2018) to regress the 3D centroids. In addition to the regression of 3D centroids, *Explanable YOLO* (Takahashi et al., 2020) used the 4-channel RGBD data directly as input to regress the 3D bounding box of pedestrians using Darknet-53 (i.e., the backbone network used in YOLOv3). When considering indoor depth images with dense pixels values, an image-based CNN approach could be considered. This is not the case when dealing with lower resolution depth images (e.g., from wearable devices). The abovementioned method focuses mainly on the positions and dimensions of the 3D pedestrians and does not take into account their orientation. However, for the collision detection systems, the orientations of the detected pedestrians play an important role - especially for predicting the movement trajectory.

The orientation of 3D objects can be represented in three ways: (1) treat the orientation estimation as a multi-class classification task by discretizing the orientations into bins (Ozuysal et al., 2009, Ghodrati et al., 2014), (2) apply direct regression (Geiger et al., 2012), and (3) in more recent years, combine both by first classifying orientation into bins and then regress the residual of orientation within a bin for refinement (e.g., F-PointNets (Qi et al., 2018)).

Frustum Pointnets (F-PointNets) (Qi et al., 2018) is a seminal work which extracts features directly from point cloud data. It first detects objects in 2D images and then extracts the foreground points using Pointnets (Qi et al., 2017). These foreground points are used to estimate the 3D bounding boxes. This method was applied to the JRDB dataset collected from a social robot (Shenoi et al., 2020). Multiple variations of F-PointNets have been proposed to improve the performance of 3D object detection. Frustum ConvNet (F-ConvNet) (Wang and Jia, 2019) is an end-to-end network, which first generates a sequence of frustums by sliding along the frustum axis with certain interval. Then, it encodes each slice of the frustum with Pointnets and encodes the sequence using a fully convolution network (FCN). A high dimensional convolution operator capturing local features enhanced with color and temporal information was proposed in (Wang et al., 2020). The work uses an early fusion approach directly on raw data from LiDAR, Camera and Radar and uses a 7D frustum representation which includes the color and precise time of the sequence. A plugin framework is designed to extract radar point cloud features efficiently. This architecture efficiently estimates the 3D bonding box along with a predicted velocity along the x and y axis.

2.3 Fusion of human pose information

Human pose estimation on 2D images are nowadays a well-developed research area. Many existing software can help researchers to easily estimate human body keypoints and generate skeletons of the people, e.g., OpenPose (Cao et al., 2019), AlphaPose (Xiu et al., 2018). Merely estimating is often not enough, hence it becomes important to make further use of this information. One of the most common application scenarios is action recognition. Handcrafted features are most frequently used. (Li et al., 2020) utilized such high level features to classify pedestrian motion state (i.e., walking and standing). Different features have been considered, e.g., positions of body keypoints reduced to neck location as origin and then normalized by the bounding box's height. In recent years, the skeletons were also encoded using Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), or Graph Convolutional Network (GCN) for action recognition (Ren et al., 2020). Furthermore, such high level features have also been used for water level estimation when people are submerged in flood water, e.g., positions of body points (Bischke et al., 2019) and distance between certain keypoints (Feng et al., 2020).

In addition to the applications above, human pose can also be

used as a clue to better estimate the orientation of the pedestrians. Pedestrians in different orientations show a different appearance. However, this information is rarely considered in the current research.

Accurate calibration between image and 3D information is important to fuse semantic information from images as in (Vora et al., 2020). When dealing with a loosely calibrated commercial-grade devices (e.g., a wearable) it would be advantages to consider higher level image feature information . In this work, we propose to introduce the 2D pose information to the 3D pedestrian detection task. This is also a very early investigation on the data collected from wearable devices. The data from such device deserve special treatment because of their typically low sampling rates and weak computing power.

3. DATA

Augmented Reality (AR) devices which can render 3D content in the real world are capable of sensing the surrounding environment. In this research, the Hololens 2 is used as a data capture device by exploiting its raw sensor streams. The wearable mixed reality device (Figure 2) offers 2D and 3D data describing its surrounding environment, which have been made available with the release of the Research Mode API (Ungureanu et al., 2020).

The far-depth sensor, which runs at lower frames of 1 to 5 fps, can be used to create 3D maps. The RGB streams from the Hololens PV (PhotoVideo) camera, along with the above-mentioned depth sensor streams, can be captured, stored, and downloaded to create RGBD datasets. This has been used to create ego pedestrian detection dataset.

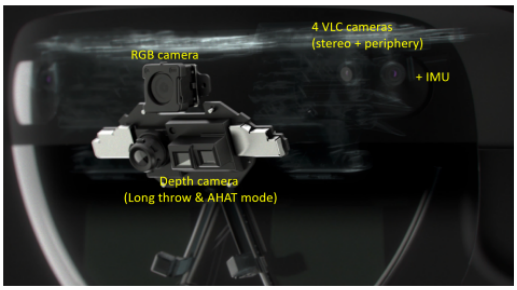

Figure 2. The Hololens 2 research mode sensors including RGB and far-depth cameras (Ungureanu et al., 2020)

3.1 The Simulated Shared Space (SSS) Dataset

An indoor experiment was conducted with a stationary Hololens user watching an open space with social interactions and pedestrian motion. In the experiment, the device captured the RGBD data of the scene. The scene recording included pedestrian encounter, dispersal, and random walk in the space of dimensions 7m x 7m.

The dataset hereby called the Simulated Shared Space dataset (SSS Dataset) records the ego view of a pedestrian wearing the sensing aid. We hypothesize that this dataset is suitable for pedestrian detection research for wearable mobility aids. The complete dataset includes 430 pedestrian frames, with each frame containing a maximum of three pedestrians. The pedestrians perform arbitrary movements and interactions in all directions within the Field of View (FoV) of the device during the experiment.

For a given RGB and its corresponding depth image, using the known calibration between the sensors, RGB information can be projected to the depth information and vice versa, exploiting the multi-sensor information for multimodal detection-based approaches.

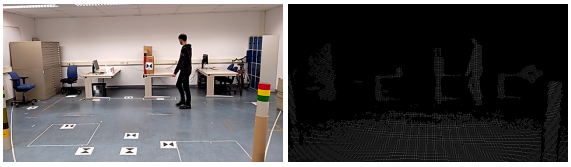

Figure 3. The RGB (left) and depth (right) data captured by ego pedestrian in the interaction space.

3.2 Annotation

The captured dataset was annotated using a semi-automated annotation procedure. As the 2D person detection in images is a well-studied problem, we used the well-known person detection, YOLO (Farhadi and Redmon, 2018), as they perform well in most cases. To obtain the 3D bounding box, region growing on the projected depth points of 2D image detection followed by box estimation was used to create ground truth automatically. This step was followed by a 3D orientation and manual error correction of pedestrians using LabelCloud (Sager et al., 2021). A total of 855 pedestrian instances (3D bounding boxes) were annotated using the above mentioned procedure with the pedestrian orientation distribution shown in Figure 4 for the complete dataset. Orientation 0 indicates that people are walking from left to right from the Hololens perspective (Figure 3).

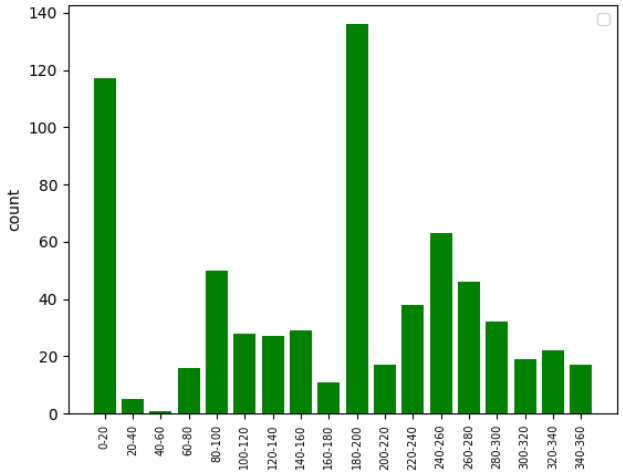

Figure 4. Orientation distribution for SSS dataset distributed into 18 bins with a bin size 20°.

4. PEDESTRIAN POSE ENHANCED FRUSTUM POINTNETS (PPEF-POINTNETS)

Given a dataset from a body worn sensor system, the aim of our work is to output 3D bounding boxes and detect people in the RGBD frames. Our framework (Figure 1) aims at improving the 3D pedestrian detection pipeline F-PointNets by including 2D human pose as additional features.

4.1 2D Pose and Hand Crafted Features

Pedestrians with different orientations will be projected as different shapes on a 2D image. This difference can provide additional clues to estimate the orientation of a pedestrian. Although this information is implicitly encoded in the local deep feature of the detected 2d pedestrians, it is not sensitive to small orientation changes. In order to explicitly encode and utilize this information, we use 2D human pose estimation and incorporate this information into the learning process.

OpenPose (Cao et al., 2019) for example, is a state-of-the-art 2D pose detection framework that can identify 25 landmark points(keypoints) of the human pose skeleton using the body-25[1] model. Once the 2D keypoints of different body parts are extracted with the framework, small variations of pixel distances between keypoints can well represent significant 3D pose information.

Even when the model detects 25 landmark points of the human body, not every keypoint can contribute to 3D pedestrian detection. For example, facial keypoints would be less accurately detected at a distance.

Hence in our work we focus along a few dominant keypoints (Figure 5) including: the shoulder (keypoint 2, 5), hip (keypoint 9, 12), knee (keypoint 10, 13), ankle (keypoint 11, 14), elbow (keypoint 3, 6), wrist (keypoint 4, 7) and neck (keypoint 1). We ignore the keypoints of the face because the distances between the keypoints are relatively short and demonstrate only little information for orientation estimation.

For the above-mentioned dominant keypoints and with an iterative feature selection and representation, we hypothesize that handcrafted pose features F_{pose} could improve the feature representation of pedestrians in the F-PointNets (Qi et al., 2018) framework. As pedestrians would appear in different distances from the camera when captured from an image source, these handcrafted features should be less sensitive to scale changes. A scale factor (SF) is introduced as a solution to this problem in our framework. SF is the distance between the shoulder and hip joints, which is used for feature normalisation:

$$SF = |Joint_{hip} - Joint_{shoulder}| \tag{1}$$

We have developed the following feature representations to be combined with the deep features from F-PointNet:

Distance Ratio (DR) : From the given set of 13 keypoint features 2D pose features, the euclidean distance between the shoulder joints and hip joints were calculated.

$$S_N = \frac{|Keypoint_{(5)} - Keypoint_{(2)}|}{SF} \tag{2}$$

$$H_N = \frac{|Keypoint_{(12)} - Keypoint_{(9)}|}{SF} \tag{3}$$

Both distances were normalized with the scale factor SF as in Equation (1). They were then applied to the network as pose features F_{pose}.

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose

$$F_{pose} = \{S_N, H_N\}$$

Optimised Distance Ratio (ODR) : The distance ratios calculated in the previous step are mostly values in the range between 0 and 1 for standing pedestrians. Smaller values correspond to people facing the camera sideways, larger values correspond to people facing the camera frontally or from behind. Very small differences in pedestrian orientations can not be well represented by simply calculating these distance ratios. Therefore we applied a negative log transformation to these ratios to exaggerate the small orientation differences to better encode the orientation information.

$$F_{pose} = \{-log(S_N), -log(H_N)\}$$

Optimised Distance with Keypoint Position and Distances (ODPD) : Inspired from the recent works in applying high-level 2D pose features (Li et al., 2020), along with the Optimised Distance Ratio, we include the normalised position and distance for all other joints, as depicted in Figure 5.

For the normalised position (N_p), as in (Li et al., 2020), the coordinates are first translated to a coordinate system with the neck as the origin. The position of the keypoints of the arms (2-7) and legs (9-14) are normalised with the SF.

Normalised distance (N_d) computes the Euclidean distance of the keypoints on the arms and legs. For the legs, four distance features would include the distance between left hip and left knee, left knee and left ankle, and the same for the right legs. While for the arms, the four features include distances between the left elbow and left shoulder, left elbow and left wrist, and corresponding features from the right arm. In total, 8 features would be normalised by the SF.

$$F_{pose} = \{-log(S_N), -log(H_N), N_p, N_d\}$$

4.2 Proposed PPEF-PointNets Pipeline

The handcrafted features F_{pose}, are introduced into the framework F-PointNets (Qi et al., 2018) by adding an off-the-shelf 2D pose detector and feature selection to the existing pipeline when detecting pedestrians. In this section, we explain our PPEF-PointNets and how we fuse the additional 2D pose to it.

A raw point cloud is obtained from RGBD scans using calibration data and depth re-projection. The raw point cloud and RGB images are inputs to the network.

Firstly, 2D pedestrians are detected in the RGB images using deep learning based 2D detections (e.g., YOLO). The 2D bounding box in RGB images is geometrically extruded to extract the corresponding frustum point cloud containing points from the point cloud that lie inside the 2D box when projected into the image plane, following the steps for frustum proposal generation as in F-PointNets.

Secondly, the RGB images are also passed to the pose detector followed by the handcrafted feature extraction F_{pose} as described in the previous section.

Methods	$AP_{0.3}$	$AP_{0.5}$	$AP_{0.7}$	AOS
Baseline F-PointNets	**0.8910**	0.4957	0.0108	0.6596
PPEF-PointNets (Distance Ratio)	0.8770	0.5004	0.0303	0.5878
PPEF-PointNets (Optimized Distance Ratio)	0.8688	**0.6470**	**0.0660**	**0.7477**
PPEF-PointNets (Optimised Distance Ratio with Keypoint Position and Distances)	0.8093	0.6358	0.0587	0.6599

Table 1. PPEF-PointNets with alternative feature selection using high level pose information

Figure 5. Pedestrian poses as detected by OpenPose (Cao et al., 2019) with the dominant keypoint features considered in our work

Thirdly, the instance segmentation module as proposed in F-PointNets applies PointNets (Qi et al., 2017) to all points contained inside the entire proposed frustum to extract features. As the features pass from the segmentation module to the amodal box estimation modules, points are transformed from the camera coordinate system to the local object coordinate system.

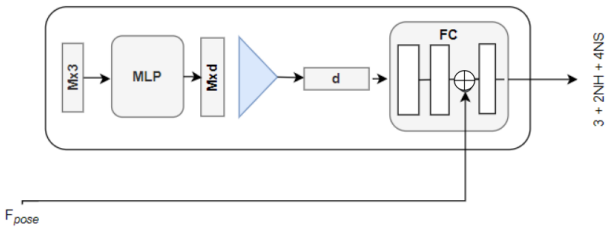

Figure 6. Bounding box estimation module with pose fusion

The 3D amodal bounding box estimation module uses these extracted point features along with the F_{pose} and applies a T-Net (Figure 6) to infer the coordinates of the 3D bounding box of the object. The loss functions used in the network are the same as proposed in F-PointNets.

5. EXPERIMENTS

With the proposed method in Section 4, we experimented with the different strategies of pose fusion in PPEF-PointNets with our SSS dataset collected with wearable sensors.

For the experimental evaluation, we have used state-of-the-art pre-trained models. While image detection using YOLOv3 pre-trained on COCO was used to detect pedestrians in the frustum proposal step, OpenPose was used to extract poses from the SSS Dataset. As the pose estimation with this multidetector works by detecting poses for multiple pedestrians occupying a single frame in one shot, a post-processing step was included to map the 2D bounding box detections to their corresponding 2D poses. With the detected poses, the handcrafted features *Distance Ratio*,*Optimised Distance Ratio* and *Optimised Distance with Keypoint Position and Distances* were computed. Hence each pedestrian for an RGB image in the SSS Dataset is characterised by a 2D bounding box and hand crafted features from the pose.

To test our network, we train our PPEF-PointNets with the three different feature sets as introduced in Section 4.1. We train for 150 epochs with a batch size of 32 and a learning rate of 0.001. The training was completed on a Nvidia 1080Ti GPU machine with the dataset randomly split into training (80%) and test sets (20%).

For comparison, we trained a F-PointNets v1 model with the SSS Dataset. This model served as the baseline for the follow-up comparison. The performance of the network was measured with the AP and AOS as used in the KITTI benchmark (Geiger et al., 2012) to indicate whether the pose feature fusion was beneficial for pedestrian 3D object detection. AP is the Average Precision often used for evaluating object detectors and AOS is the Average Orientation Similarity proposed in KITTI for evaluating 3D orientations. The quantitative evaluation is summarized in Table 1.

As can be seen from the results, the optimised distance features improves (except IoU=0.3) over the baseline in overall 3D pedestrian detection performance. In contrast to the expectations and also as in other works (Yu et al., 2019) where the position and distance of keypoints were used, adding them in ODPD did not show further improvements. This may be due to the fact that such features do not cope well with changes in perspective and different poses of people.

In further, the AP and AOS at different IoU thresholds are visualized in Figure 8. The model using 2D pose information achieved better performance for almost all the IoU thresholds.

To further evaluate the improvement in accuracy, an IoU of 0.1 was set, and all the true positives detected in Optimised Distance (ODR) were compared with the baseline. It can be noted that 63% of the True Positives detected show improved 3D IoU when the pose is added, with a mean improvement of 13%. We consider this as clear evidences for the benefits of integrating 2D handcrafted pose features, as done by our approach. Leveraging reliable 2D pose estimates yields a higher performance for 3D detection. Furthermore, we also achieve a lower error for the orientation estimation.

Qualitative results are presented in Figure 7. The results of the model with the best performance are compared with the results of the baseline approach. From the visualization in 3D, the

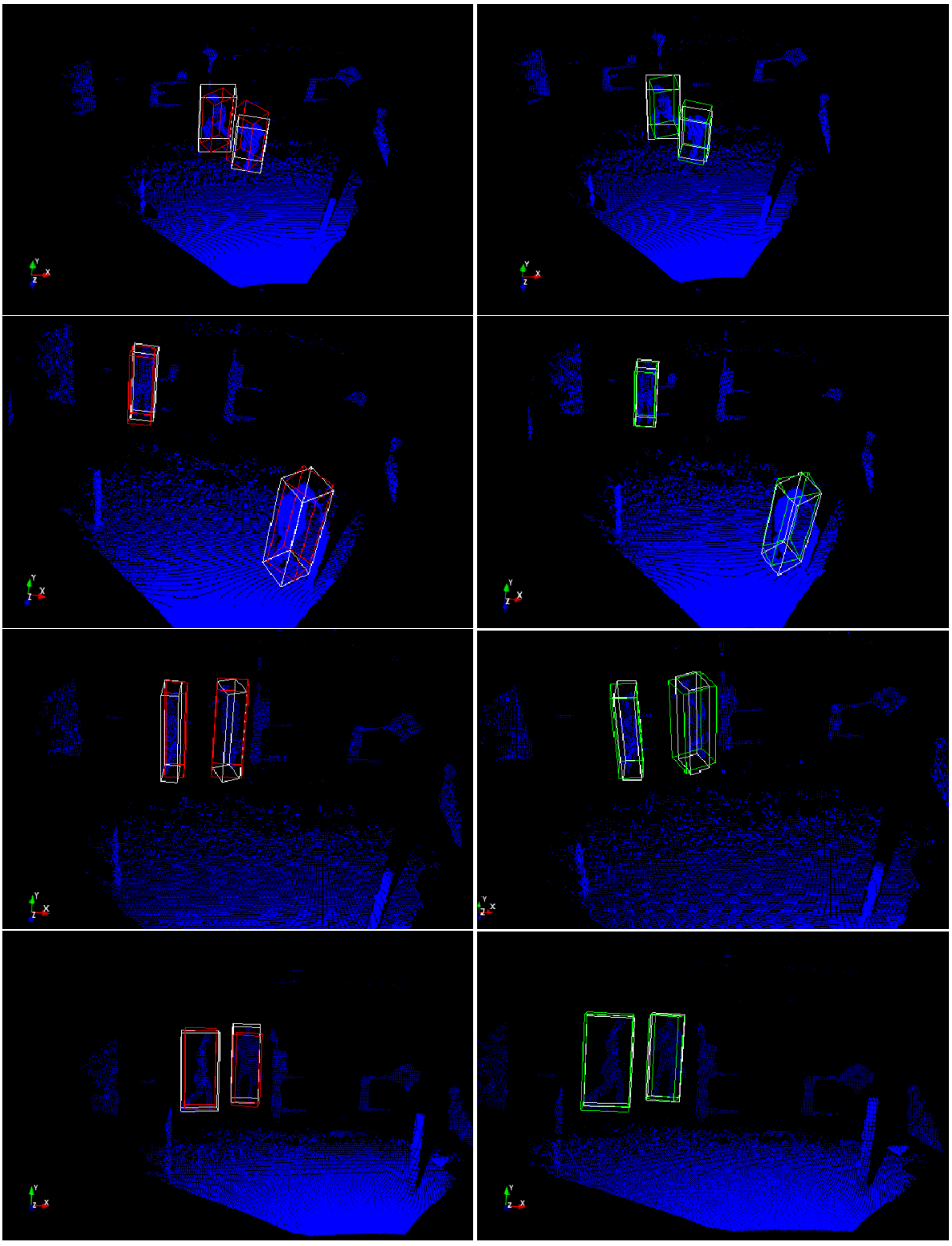

Figure 7. Qualitative comparison of pedestrian 3D detection results using baseline (red bounding boxes on the left) and our proposed approach using ODR features (green bounding boxes on the right). The white bounding boxes are the manually annotated ground truth.

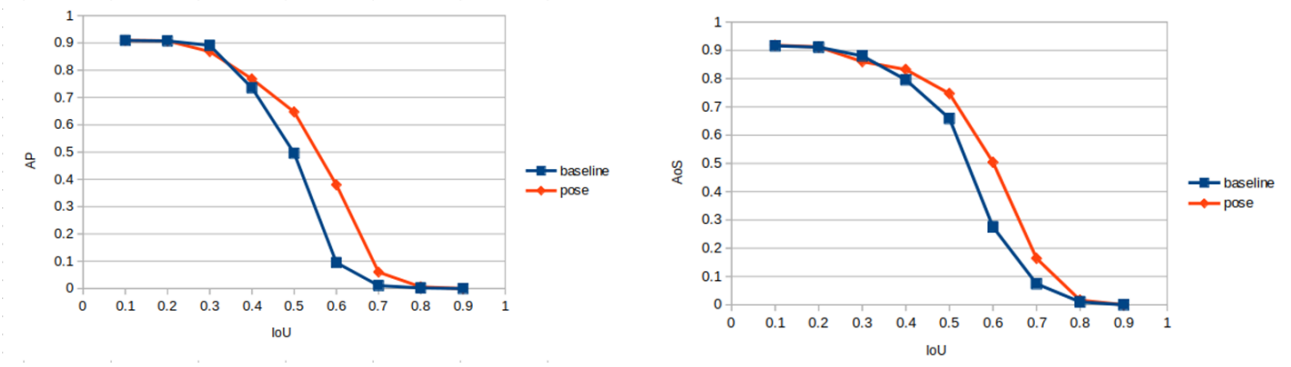

Figure 8. The AP and AOS for different values of IoU threshold for ODR compared against Baseline F-PointNets

Figure 9. Qualitative comparison of pedestrian 3D detection results in bird eye view using baseline (red bounding boxes) and our proposed approach using ODR features (green bounding boxes). The black bounding boxes are the manually annotated ground truth. The edge in cyan color indicate the frontal orientation of the pedestrian.

orientations of the 3D bounding boxes are better estimated (according to the examples in the first three rows). The proposed model also performs better for estimating the pedestrian dimensions (see the last row).

In order to present the localization performance of pedestrian detection, examples of 3D pedestrian detection are visualized in bird's eye views as in Figure 9. A better performance has been achieved by the proposed method.

6. CONCLUSIONS AND OUTLOOK

In this paper, we realised 3D pedestrian detection on data collected from wearable sensors. F-PointNet was fused with extra high-level 2D human pose information via experimenting with three types of handcrafted features. The newly proposed pipeline demonstrated a better performance compared to the original F-PointNet.

However, our framework has only been tested with a dataset collected from wearable Hololens device with less complicated indoor scenarios. Investigation of the performance of our approach on newly published indoor pedestrian detection datasets (Shenoi et al., 2020) could be a direction for future work. Also as the 3D detection system is intended for AR device, its real time performance and efficiency has to be evaluated in a next step.

While detection is an important component in collision detection, realising other components of a detection pipeline (e.g., 3D tracking) by including the improved orientation estimates and motion prediction (Cheng et al., 2020) need to be addressed in the near future. Combining this collision detection system with a 3D interface via a wearable device could leverage the power of perception and visualisation and hence controlling the walking behaviour of pedestrians in shared spaces.

ACKNOWLEDGEMENTS

The authors cordially thank the funding provided by DAAD and participants of DFG Training Group 1931, SocialCars on the directions used in the work.

REFERENCES

Bischke, B., Brugman, S., Helber, P., 2019. Flood severity estimation from online news images and multi-temporal satellite images using deep neural networks. *MediaEval*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y., 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 172–186.

Cheng, H., Liao, W., Yang, M. Y., Sester, M., Rosenhahn, B., 2020. Mcenet: Multi-context encoder network for homogeneous agent trajectory prediction in mixed traffic. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 1–8.

Farhadi, A., Redmon, J., 2018. Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition*, Springer Berlin/Heidelberg, Germany, 1804–2767.

Feng, Y., Brenner, C., Sester, M., 2020. Flood severity mapping from Volunteered Geographic Information by interpreting water level from images containing people: A case study of Hurricane Harvey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169, 301–319.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 3354–3361.

Ghodrati, A., Pedersoli, M., Tuytelaars, T., 2014. Is 2d information enough for viewpoint estimation? *Proceedings BMVC 2014*, 1–12.

Hamilton-Baillie, B., 2008. Shared space: Reconciling people, places and traffic. *Built environment*, 34(2), 161–181.

Ito, T., Kamata, M., 2013. Autonomous locomotion based on interpersonal contexts of pedestrian areas for intelligent powered wheelchair. *International Conference on Human Interface and the Management of Information*, Springer, 480–489.

Javid, B., Seneviratne, P. N., 1991. Applying conflict technique to pedestrian safety. *ITE journal*.

Kamalasanan, V., Sester, M., 2020. Living with rules: An ar approach. *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, IEEE, 213–216.

Kollmitz, M., Eitel, A., Vasquez, A., Burgard, W., 2019. Deep 3D perception of people and their mobility aids. *Robotics and Autonomous Systems*, 114, 29–40.

Li, F., Fan, S., Chen, P., Li, X., 2020. Pedestrian motion state estimation from 2d pose. *2020 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 1682–1687.

Linder, T., Pfeiffer, K. Y., Vaskevicius, N., Schirmer, R., Arras, K. O., 2020. Accurate detection and 3d localization of humans using a novel yolo-based rgb-d fusion approach and synthetic training data. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 1000–1006.

Ozuysal, M., Lepetit, V., Fua, P., 2009. Pose estimation for category specific multiview object localization. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 778–785.

Poggi, M., Nanni, L., Mattoccia, S., 2015. Crosswalk recognition through point-cloud processing and deep-learning suited to a wearable mobility aid for the visually impaired. *International Conference on Image Analysis and Processing*, Springer, 282–289.

Qi, C. R., Liu, W., Wu, C., Su, H., Guibas, L. J., 2018. Frustum pointnets for 3d object detection from rgb-d data. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 918–927.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Ren, B., Liu, M., Ding, R., Liu, H., 2020. A survey on 3d skeleton-based action recognition using learning method. *arXiv preprint arXiv:2002.05907*.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 91–99.

Sager, C., Zschech, P., Kühl, N., 2021. labelcloud: A lightweight domain-independent labeling tool for 3d object detection in point clouds.

Shenoi, A., Patel, M., Gwak, J., Goebel, P., Sadeghian, A., Rezatofighi, H., Martín-Martín, R., Savarese, S., 2020. Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 10335–10342.

Takahashi, M., Ji, Y., Umeda, K., Moro, A., 2020. Expandable yolo: 3d object detection from rgb-d images. *2020 21st International Conference on Research and Education in Mechatronics (REM)*, IEEE, 1–5.

Ungureanu, D., Bogo, F., Galliani, S., Sama, P., Meekhof, C., Stühmer, J., Cashman, T. J., Tekin, B., Schönberger, J. L., Olszta, P. et al., 2020. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*.

Vora, S., Lang, A. H., Helou, B., Beijbom, O., 2020. Pointpainting: Sequential fusion for 3d object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4604–4612.

Wang, J., Wood, Z., Worboys, M., 2016. Conflict in pedestrian networks. *Geospatial Data in a Changing World*, Springer, 261–278.

Wang, L., Chen, T., Anklam, C., Goldluecke, B., 2020. High dimensional frustum pointnet for 3d object detection from camera, lidar, and radar. *2020 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 1621–1628.

Wang, Z., Jia, K., 2019. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 1742–1749.

Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C., 2018. Pose Flow: Efficient online pose tracking. *BMVC*.

Yu, D., Xiong, H., Xu, Q., Wang, J., Li, K., 2019. Continuous pedestrian orientation estimation using human keypoints. *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 1–5.