Body and Head Orientation Estimation with Privacy Preserving LiDAR Sensors

Onur N. Tepencelik, Wenchuan Wei, Leanne Chukoskie†, Pamela C. Cosman, Sujit Dey
Dept. of Electrical and Computer Engineering, †Qualcomm Institute
University of California San Diego
San Diego, USA
{otepence, w8wei, lchukoskie, pcosman, dey} @ucsd.edu

***Abstract*—Body and head orientation estimation is important in many applications such as pedestrian protection, movement prediction, robotics, and behavioral analysis. In this paper, we propose a system that uses privacy preserving LiDAR sensors to estimate body and head orientations of people, with a motivation of providing guidance feedback to individuals who face non-verbal social communication challenges in workplace settings, such as some individuals with Autism Spectrum Disorder (ASD). For example, people who tend to look away from a speaker could be coached on the importance of periodically making eye contact and showing overt attention, or could be discreetly provided with real-time feedback based on present behavior. We developed models for body and head orientation estimation, using low-resolution point cloud data from two LiDAR sensors. The body orientation estimation model uses an ellipse fitting method while the head orientation estimation model is a pipeline of geometric feature extraction and neural network-based regression. Compared with other body and head orientation estimation systems using RGB cameras, our proposed system uses LiDAR sensors to preserve user privacy, while achieving comparable accuracy. To the best of our knowledge, this is the first body and head orientation estimation system using depth sensors for which the sensors do not require a specified placement in front of the subject. Our model achieves a mean absolute estimation error of 8.4 degrees for body orientation and 16.5 degrees for head orientation.**

***Index Terms*—Body orientation, head orientation, LiDAR sensor, point cloud, autism spectrum disorder**

I. Introduction

Body and head orientation estimation are fundamental challenges in computer vision, mainly investigated in the context of pedestrian protection and movement prediction [1], along with applications in robotics [2] and behavior analysis [3]. Body and head orientation and movement provide important means of nonverbal communication for fluent social interaction. Some people with social communication deficits (for example, some individuals with Autism Spectrum Disorder (ASD)) struggle to provide normative nonverbal communication cues, such as periodically making eye contact with the speaker and maintaining an overall appropriate body orientation towards them [4]. Lack of workplace-appropriate social communication skills are one reason that high-functioning young adults with ASD have high unemployment rates despite often holding college degrees, average to high IQs, and various useful skills. In this paper, we focus on the problem of body and head orientation estimation from a surveillance viewpoint, with the primary motivation of providing guidance feedback to individuals who face social communication challenges in typical workplace settings.

Most works on body and head orientation estimation use RGB cameras for their low cost [3], [5], [6], but RGB-D cameras such as Microsoft Kinect and Intel RealSense have also been used [7]–[10]. Available depth image-based models using RGB-D sensors or LiDARs seemed to be good candidates to solve our problem. However, these models require the sensor to be placed in front of the person, with specific optimal ranges for distance and height, which we will refer to as a *frontal setting*. In contrast, our system does not require the subject to appear head-on in front of the sensor. Our sensors are placed near the ceiling, looking down at about 45 degrees, and the subject can have arbitrary orientation in the conference area; we refer to this setup as a *surveillance setting*. To the best of our knowledge, there is no model for body and head orientation estimation with depth cameras or LiDAR sensors from a surveillance viewpoint. In general, surveillance settings produce low resolution data. As the subject gets more distant from the sensor, they are represented with fewer points in a point cloud or fewer pixels in an RGB image. Especially for head pose estimation, most models [9], [10] use high-resolution 3D scans of the head, taken by a sensor close to the subject. With such a setting, it is possible to capture small facial geometric details of the nose tip, eye holes, and chin, which can play a huge role for orientation estimation. While those models are successful for high-resolution data, their methods face challenges in our case, as the sensors are unobtrusively distant from the people, and it is difficult to identify small facial geometric features due to the decreasing resolution and increasing noise with distance.

Our contributions in this paper include the development of body and head orientation estimation models based on low resolution point cloud data, generated by two indoor LiDAR sensors from a surveillance viewpoint. Fig. 1 shows the system overview. The LiDAR sensors in our system are specialized for indoor use, and come with built-in functionality to detect humans in the field of view. We stitch together the point clouds from the two sensors, extract the upper bodies by cropping a cylinder-shaped point cloud around each detected person, and remove noise points. Then we extract features, fit a least squares ellipse to determine the upper body, and use a multi-layer perceptron based neural network regressor with the extracted features to estimate the head orientation.

This work was supported by the National Science Foundation under grant DUE-1928604.

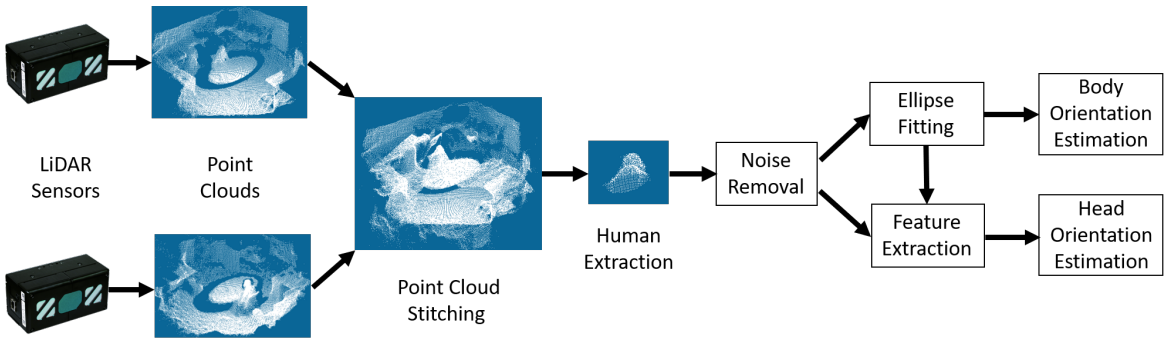

Fig. 1. System overview.

II. Related Work

Among the many RGB image-based models for body orientation estimation which are generally in the context of smart vehicles and robotics for human-robot interactions, there are a few which match our type of surveillance setting. Chen *et al.* [3] proposed a semi-supervised model on RGB images to analyze behavior and attention based on estimated body and head orientations of people waiting for luggage in an airport. The authors of [5] and [6] proposed template matching models that combine 2D images from multiple surveillance viewpoints to make 3D orientation estimations. The few works on body orientation estimation using depth sensors do not use surveillance scenarios. Shimizu *et al.* [11] proposed a body orientation estimation model which combines shape and motion information, using a LiDAR sensor mounted on a robot. The authors of [2] and [12] proposed models that use depth along with color information.

Most papers on head orientation focus on estimating the full head pose by estimating the three Euler angles, whereas we only focus on estimating a direction angle, as other dimensions are not important for our use case. As with body orientation, the majority of models in the literature use RGB images, but some use depth, and only a few consider the task in a surveillance setting. The authors of [1], [3], [13] proposed various models and surveillance settings for the head orientation estimation task using RGB cameras. Following the availability of consumer level depth cameras such as Microsoft Kinect and Intel RealSense, various models [7]–[10] targeting head orientation estimation using depth images were proposed. However, similar to the body orientation estimation models, the models listed above assume that the subject is in a frontal setting, with the depth sensor in front of the person.

To the best of our knowledge, we are the first to estimate body and head orientations using a depth sensor from a surveillance viewpoint. Existing models are either RGB image based on a surveillance setting, or depth image based on a frontal setting.

III. Methodology

A. Data Collection and Preprocessing

We developed our models using two LiDAR sensors from Hitachi Vantara [14]. We use the ToFv2 LiDAR sensor which works based on the Time-of-Flight principle [15]. For data collection, the sensors were placed on opposite ceiling corners of a small conference room, looking down on an oval table. The point clouds coming from the sensors are stitched together using rotation and translation before processing. Then we use the built-in human detection capability of the sensor software to crop a cylinder-shaped boundary around each person's point cloud. The human detection capability allows us to process each person in the environment separately at the same time, and decreases the number of points to be processed by the model. We further crop the point clouds to get the region of interest, which includes the upper body and head. For each subject, the cropping threshold for the upper body was set to be the top 27% of their height (in a seated position).

We created a dataset from 15 subjects including men and women with and without glasses and face masks, and with varying hairstyles and heights. We collected the data one subject at a time, while the subject sits in 8 different positions around the oval table in the middle of the room. Guidance arrows are placed on the table for each seat position and each orientation direction to determine the ground truth, and the point clouds are captured while each subject orients their head towards 13 predetermined angles (-90 to +90 degrees, in increments of 15 degrees). Thus, for each subject, we attempted to collect 312 point clouds, corresponding to 8 seat positions, 13 head orientation angle, and 3 repetitions. Because a few repetitions were missed or the human tracking failed, we ended with 295 point clouds per person on average. Our upper body point clouds consist of about 1800 points on average, varying between about 1500 and 2100 points per case. Our resolution is low compared to the BIWI dataset [8], also used in [9], [10], which contain around 10,000 points for just the face of a person.

Estimating body and head orientation from LiDAR data is a challenge as the sensor cannot capture the details of the small region of interest from a distance of 1 to 4 meters. Moreover, the point cloud data are noisy, especially from hair and other complex features on the face. To mitigate this problem, we apply a k-nearest neighbors-based noise removal pre-processing step, where we remove a point from the point cloud if the average distance between the point and its 10 nearest neighbors is larger than 40 millimeters.

B. Body Orientation Estimation

The body orientation estimation model is a geometric model which takes advantage of the ability to change the viewpoint of a point cloud, and uses the birds-eye view of the room. The cropped point clouds are projected onto the xy-plane (the plane parallel to the ground). We exclude the head points by removing all points within 20 centimeters of the top of the head, and calculate the 2D ellipse that best fits the projected upper body data points based on least squares error, with the long axis of the ellipse representing the body axis. We use the conic representation of an ellipse:

$$E(x, y) = ax^2 + bxy + cy^2 + dx + ey + f = 0 \qquad (1)$$

The optimal coefficients are estimated following the approach in [16]. The noise removal pre-processing step is important for this procedure to work well, as noise points that are generally on the edges may result in large squared errors and significantly disturb the best ellipse fit.

After the body axis is determined, there remains the problem of deciding which side of the ellipse is the person's front. We use the positions of the previously excluded head points with respect to the center of the ellipse, based on the fact that a person's head is almost always in front of their body axis.

C. Head Orientation Estimation

For the head orientation estimation model, simple geometric approaches were not sufficient as details of facial features are not accurately captured by the sensors. We use a pipeline of feature extraction and multi-layer perceptron-based regression. The feature extraction is done after noise removal and ellipse fitting to the upper body. The upper body ellipse divides the head point cloud into four quadrants which produce supportive features for the model, based on the point locations with respect to the body center. Fig. 2 shows a projected human point cloud, the optimal ellipse fit for body orientation estimation, and the resulting four quadrants of the head.

The features we extract are the (x, y) coordinates of the subject's centroid in the sensor coordinate system, as well as a number of features that use a subject-centric coordinate system. These features are the principal components and the basic distribution properties of the head points (mean, standard deviation, minimum and maximum coordinates), the principal components of the four quadrants of the head, the estimated nose coordinates based on the centroid of the 10 furthest projected points from the head center, and the axis lengths and orientations of a separate ellipse fitting procedure on the head points only.

We note that the cropped human point clouds vary widely across the 8 different positions around the table in our data collection setting, due to the positioning of the sensors and the corresponding occlusion of some body parts. In each position, it is likely that some features are informative while others are ineffective. For example, in some positions, the estimated nose position is almost the same regardless of how much the subject turns their head from left to right. In other positions, the principal components do not work well because the overall rotation of the head is not well represented by the point clouds. Therefore, we use a learning-based model, which is able to learn the variations based on the relative position of the subject and weight useful features, with robustness to the low resolution, high noise and positional variations.

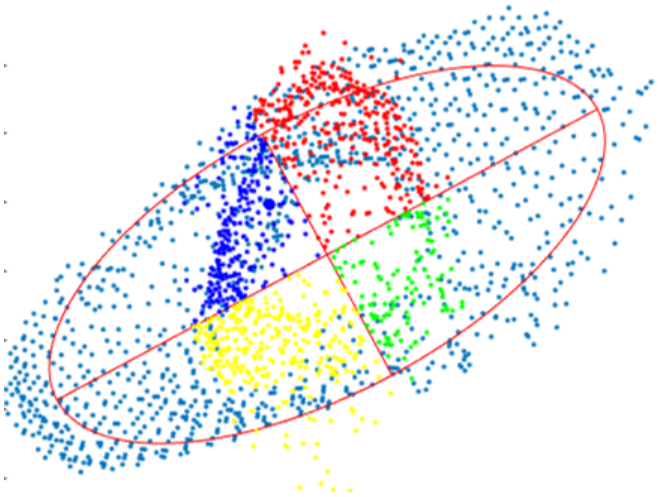

Fig. 2. Least squares ellipse fitting for body orientation estimation via the long axis of the fitted ellipse, which also determines the four quadrants of the head relative to the body. Light blue points are projected upper body points (shoulders and chest); dark blue, red, yellow and green points are projected head points, representing the four quadrants.

IV. Results

We use two metrics to evaluate performance. The first is the mean absolute error (MAE) between the model estimation and ground truth. The body orientation estimation model produces an MAE of 8.4 degrees over the whole dataset. MAEs with respect to seating position are in Table I, where the seating positions are as shown in Fig. 3. We observe that the average MAEs of the subjects sitting in positions 2 and 6 are the lowest; we surmise this low MAE arises because those subjects are squarely facing one of the sensors and have their backs squarely turned to the other sensor, so there would be large body areas with good data capture. The model struggles when large parts of the upper body are not visible to the sensors, since missing points distort the fitted ellipse.

For training the head orientation model, we use leave-one-out cross-validation, where the point clouds of each data

TABLE I
Mean Absolute Orientation Errors vs. Seating Position

Position	Body MAE	Head MAE
P1	12.21	12.69
P2	4.47	14.96
P3	8.95	19.11
P4	6.93	15.77
P5	13.49	15.01
P6	4.63	17.48
P7	8.96	22.63
P8	7.12	14.57
Average	8.37	16.49

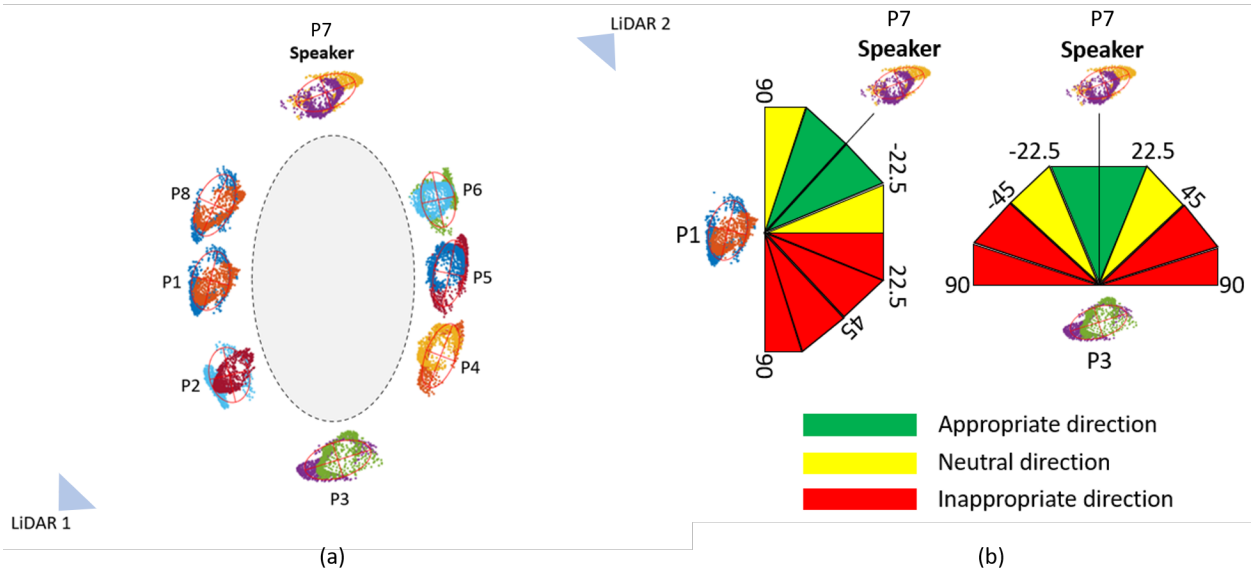

Fig. 3. (a) 8-person meeting scenario with a fixed speaker at the top who is assumed to be the main target of head orienting for the other participants. (b) Appropriateness regions of head orientations for persons in Position 1 and Position 3 with respect to the speaker.

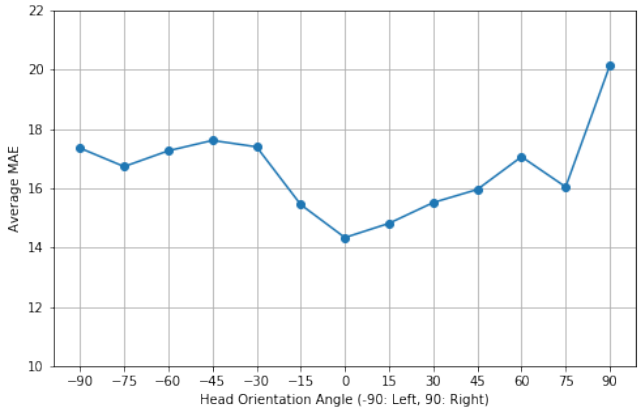

Fig. 4. Mean Absolute Error vs. Head Orientation Ground Truth

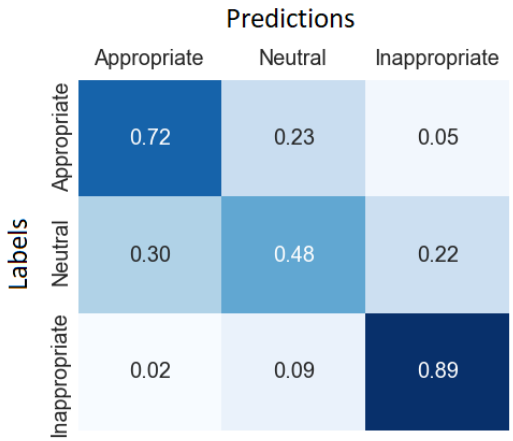

Fig. 5. Confusion matrix of appropriateness evaluation.

subject are the test set one time, and used in training otherwise. The model has an average MAE of 16.5 degrees across 15 subjects. MAEs with respect to seating position are in Table I, and MAEs with respect to ground truth angle are in Fig. 4. While some papers achieve smaller errors on head orientation estimation, they use either high-resolution 3D scans of the face when the sensor is placed right in front of the person [7]–[10], or an RGB camera [17].

The second evaluation metric aims at a social communication context, since one eventual goal of this work is to coach people with ASD who struggle with nonverbal communication in a conversation or a meeting. In a speaker-listener scenario, we examine the subject's head orientation towards the speaker. Fig. 3 shows an example of an 8-person meeting scenario with regions of appropriateness and non-appropriateness. From a virtual coaching perspective, if the body or head orientation of the subject is in the red region for some length of time, a real-time alert could remind the subject to maintain nonverbal contact with the speaker. The green region means that the subject is maintaining appropriate body and head orientations, and the yellow region might help with coaching decisions based on more temporal information from previous or following frames. In our experiments, we set the directional range of appropriateness to be within 45 degrees, meaning that the limits for the green region would be -22.5 and 22.5 degrees. We assumed that if the subject is off from the speaker by at least 45 degrees (for some length of time), the situation is inappropriate. Orientations between 22.5-45 degrees are neutral and action could be taken based on temporal analysis. For example, the neutral situation could be taken to be inappropriate if it occurs for more than 30 seconds. The temporal analysis is beyond the scope of this paper; we examined single frame estimation results from the

whole dataset. Fig. 5 shows the confusion matrix results from our experiments. The results are based on 8 different scenarios where we assumed that the speaker is in one of the 8 positions around the table and the listeners are in the remaining 7 positions.

The number of false alarms (raising an alarm when the subject is keeping appropriate orientations) is fairly low. Our model is accurate in distinguishing between appropriate and inappropriate situations, and most of the false predictions fall into the neutral regions which will not cause huge errors in our future application. In conjunction with behavioral coaching experts, our next step will be the development of the virtual coaching model, involving analysis of a sequence of point clouds over time, and addressing, among other issues, when and how to provide coaching feedback.

V. Conclusion

In this paper, we propose models for body and head orientation estimation that work with low resolution point clouds generated by two LiDAR sensors. The main motivation of this work is to create a privacy-preserving system that could be used as a virtual coach for people, such as some individuals with ASD, who struggle nonverbal communication including maintaining appropriate body and head poses during conversations or meetings. To achieve this, we created a surveillance scenario with LiDAR sensors placed near the ceiling of a conference room and developed novel models that can estimate the body and head orientations of the subjects from the low-resolution point cloud data. Our body and head orientation estimation models produce average error rate of 8.4 and 16.5 degrees, respectively. Our results are comparable to results in the literature, although our models work with low-resolution and noisy point clouds and without color information. The proposed system is able to distinguish between appropriate and inappropriate body and head orientations.

The proposed body and head orientation estimation models can be used in various applications. We plan to extend our models to become on component of virtual coaching to high-functioning individuals with ASD who are seeking jobs, to integrate them to workplaces. With the recent advances of LiDAR sensors and declining costs, they are likely to become more prevalent [18], [19] in stores and workplaces, thanks to their privacy preservation aspect. When developing a virtual behavioral coaching system, a main consideration is whether the system will provide real-time alerts as inappropriate behaviors occur, and avoid alerts for appropriate behavior. While the computational power is available to make estimations and provide guidance in real-time with this model, the system could also be used as a non-real-time social behavior analysis tool for people with ASD by providing feedback based on their overall display of attentiveness and non-verbal communication performance in a conference. Other considerations include the medium (smart watch notifications, vibrations, audio etc.) and the content of the alerts or feedback.

Although outside of our immediate focus, the system could also be modified for outdoor use using appropriate LiDAR products, as assistants to security cameras. Various applications such as crowd analysis or protection of pedestrians crossing streets could benefit from this type of system.

References

[1] E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 2292–2297.

[2] B. Lewandowski, D. Seichter, T. Wengefeld, L. Pfennig, H. Drumm, and H.-M. Gross, "Deep orientation: Fast and robust upper body orientation estimation for mobile robotic applications." in *IROS*, 2019, pp. 441–448.

[3] C. Chen and J.-M. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1544–1551.

[4] V. H. Bal, S.-H. Kim, M. Fok, and C. Lord, "Autism spectrum disorder symptoms from ages 2 to 19 years: Implications for diagnosing adolescents and young adults," *Autism Research*, vol. 12, no. 1, pp. 89–99, 2019.

[5] L. Chen, G. Panin, and A. Knoll, "Human body orientation estimation in multiview scenarios," in *International Symposium on Visual Computing*. Springer, 2012, pp. 499–508.

[6] M. C. Liem and D. M. Gavrila, "Person appearance modeling and orientation estimation using spherical harmonics," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[7] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3D head pose estimation using the Kinect," in *2011 International conference on wireless communications and signal processing (WCSP)*. IEEE, 2011, pp. 1–4.

[8] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Joint pattern recognition symposium*. Springer, 2011, pp. 101–110.

[9] P. Padeleris, X. Zabulis, and A. A. Argyros, "Head pose estimation on depth data based on particle swarm optimization," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 42–49.

[10] C. Papazov, T. K. Marks, and M. Jones, "Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4722–4730.

[11] M. Shimizu, K. Koide, I. Ardiyanto, J. Miura, and S. Oishi, "LIDAR-based body orientation estimation by integrating shape and motion information," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2016, pp. 1948–1953.

[12] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li, "Accurate estimation of human body orientation from RGB-D sensors," *IEEE Transactions on cybernetics*, vol. 43, no. 5, pp. 1442–1452, 2013.

[13] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," in *International evaluation workshop on classification of events, activities and relationships*. Springer, 2006, pp. 299–304.

[14] "Hitachi vantara 3D LiDAR," https://www.hitachivantara.com/en-us/products/video-intelligence/devices/3d-lidar-sensor.html, [Online].

[15] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *IEEE Journal of quantum electronics*, vol. 37, no. 3, pp. 390–397, 2001.

[16] F. L. Bookstein, "Fitting conic sections to scattered data," *Computer graphics and image processing*, vol. 9, no. 1, pp. 56–71, 1979.

[17] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *CVPR 2011*. IEEE, 2011, pp. 617–624.

[18] C. Sebastian, B. Boom, E. Bondarev *et al.*, "LiDAR assisted large-scale privacy protection in street view cycloramas," *Electronic Imaging*, vol. 2019, no. 11, pp. 281–1, 2019.

[19] A. Günter, S. Böker, M. König, and M. Hoffmann, "Privacy-preserving people detection enabled by solid state LiDAR," in *2020 16th International Conference on Intelligent Environments (IE)*. IEEE, 2020, pp. 1–4.