Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Body and Head Orientation Estimation from Low-Resolution Point Clouds in Surveillance Settings

ONUR N. TEPENCELIK, WENCHUAN WEI, PAMELA C. COSMAN, (Fellow, IEEE), SUJIT DEY, (Fellow, IEEE)

Electrical and Computer Engineering, UC San Diego, La Jolla, CA 92093, USA. (e-mail: otepence, w8wei, pcosman, sdey @ucsd.edu)

Corresponding author: Onur N. Tepencelik (e-mail: otepence@ucsd.edu).

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

ABSTRACT We propose a system that estimates people's body and head orientations using low-resolution point cloud data from two LiDAR sensors. Our models make accurate estimations in real-world conversation settings where subjects move naturally with varying head and body poses, while seated around a table. The body orientation estimation model uses ellipse fitting while the head orientation estimation model combines geometric feature extraction with an ensemble of neural network regressors. Our models achieve a mean absolute estimation error of 5.2 degrees for body orientation and 13.7 degrees for head orientation. Compared to other body/head orientation estimation systems that use RGB cameras, our proposed system uses LiDAR sensors to preserve user privacy, while achieving comparable accuracy. Unlike other body/head orientation estimation systems, our sensors do not require a specified close-range placement in front of the subject, enabling estimation from a surveillance viewpoint which produces low-resolution data. This work is the first to attempt head orientation estimation using point clouds in a low-resolution surveillance setting. We compare our model to two state-of-the-art head orientation estimation models that are designed for high-resolution point clouds, which yield higher estimation errors on our low-resolution dataset. We also present an application of head orientation estimation by quantifying behavioral differences between neurotypical and autistic individuals in triadic (three-way) conversations. Significance tests show that autistic individuals display significantly different behavior compared to neurotypical individuals in distributing attention between conversational parties, suggesting that the approach could be a component of a behavioral analysis or coaching system.

INDEX TERMS Autism spectrum disorder, body orientation, head orientation, LiDAR sensor, point cloud, triadic conversation, triadic interaction

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]. Most work on body and head orientation estimation uses RGB cameras for their low cost and prevalence [3], [4], but more expensive RGB-D cameras such as Microsoft Kinect and Intel RealSense have also been used [5], [6]. However, use of RGB cameras raises privacy concerns in many cases. Studies suggest that people's concerns over privacy have been increasing, with privacy protection mechanisms getting more attention [7], [8]. We propose a system that uses point cloud data from LiDAR sensors to estimate body and head orientations while protecting user privacy. While depth maps also preserve privacy, most common depth sensors are RGB-D that record color information as well, whereas LiDAR solely outputs depth, making it a more privacy-safe device [9]. There has been increased adoption of LiDAR sensors with declining costs [10]. Many recent projects in different fields such as healthcare [9], [11], security, and surveillance [12], [13], have adopted LiDAR sensors over privacy invading alternatives. With recent advances in LiDAR technology and big data management systems that enable data scalability [14], they are likely to become more prevalent in stores, workplaces, and hospitals.

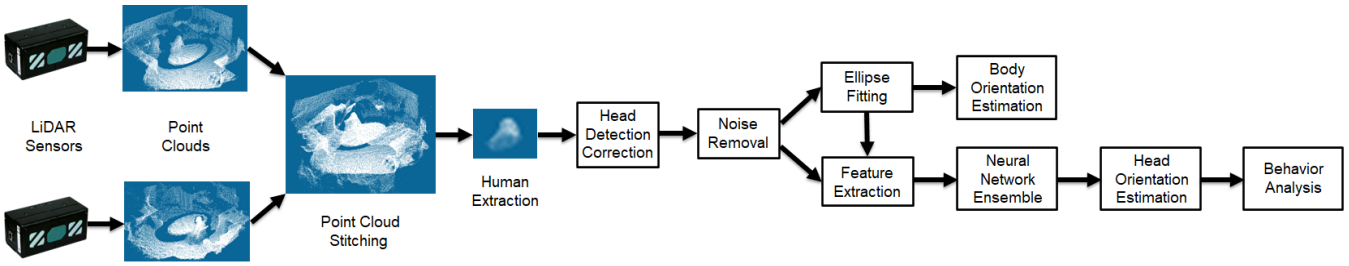

FIGURE 1. System overview.

Available depth image-based models using RGB-D sensors or LiDAR seemed to be good candidates for our need to estimate body and head orientation. However, these models require the sensor to be placed in front of the person, with specific optimal ranges for distance and height, which we 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 as a *surveillance setting*. Our models for body and head orientation estimation with LiDAR sensors are the first that permit a surveillance viewpoint.

In general, surveillance settings produce low-resolution data; a subject farther from the sensor is represented with fewer points in a point cloud or fewer pixels in an RGB image. Especially for head pose estimation, most models [15], [16] 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, they face challenges in our case, as our sensors are unobtrusively distant from the people, and it is difficult to identify small facial geometry features due to the decreasing resolution and increasing noise with distance.

Being able to work with low-resolution data is essential for models targeting a surveillance setting. For example, pedestrian protection applications that aim to detect pedestrians and predict their movements from a surveillance viewpoint or from a sensor mounted on a vehicle would benefit from a system that enables estimations from low-resolution data [1], [17]. Similarly, Chen *et al.* [3] proposed a head and body orientation estimation model to analyze and predict behaviors in public spaces such as airports, which could be useful for public safety. Various other studies proposed leveraging head orientation estimation systems for attention and interaction modeling, for instance in museums to analyze which artworks are getting more interest [18] or in shopping centers to gauge which products are attracting more customers [19], or in work environments to analyze social interactions [20].

In this paper, we also present an application of our head and body orientation estimation models. Multiple studies [21], [22] have shown that head orientation is a good indicator of visual focus of attention, without the need to estimate gaze orientation. Body and head orientation and movement provide important means of nonverbal communication for fluent social interaction. Individuals with social communication differences (for example, some individuals with Autism Spectrum Disorder (ASD)) might not regularly provide normative nonverbal communication cues, such as periodically making eye contact with a speaker and maintaining a body orientation generally towards them [23]. Differences from society's workplace communication norms are one reason that high-functioning young adults with ASD have high unemployment rates [24] despite often holding college degrees, average to high IQs, and various useful skills. Furthermore, it was found that many autistic people were terminated from jobs due to communication differences [25].

To analyze behaviors related to body and head orientation, we use a triadic (three-way) conversation setting with two *interviewers* and one *subject* sitting around an oval conference table. Triadic conversations are common in professional and social settings and they are harder to navigate compared to dyadic (two-way) interactions [26]. While some autistic individuals may find it challenging to show attention in a dyadic interaction with potential distractions such as objects, a triadic interaction involves an additional person who is a part of the conversation and may require attention. Adjusting body and head orientation in triadic settings is important to engage with both of the other people and make everyone feel included in the conversation [27].

In triadic interactions, some individuals with ASD tend to fixate on one person while ignoring the other for some time [28], a non-normative distribution of attention which could be seen as non-inclusive or socially inappropriate. Other neurodivergent behaviors commonly displayed by autistic individuals include not making eye contact with any of the interviewers while speaking, or not paying attention to a speaker while listening [29], [30]. Our body and head orientation estimation system can quantify such behavioral differences between autistic and neurotypical individuals. We plan to extend our system to provide coaching and feedback to autistic individuals, imitating the coaching advice of a professional behavioral coach, with the motivation of supporting autistic individuals in practicing conversational engagement skills in preparation for job interviews and workplace communications [31], [32].

Our main contribution in this paper is the development of novel body and head orientation estimation models specifically designed to work with low-resolution point cloud data, generated by two indoor LiDAR sensors from a surveillance viewpoint. Fig. 1 shows the system overview. A preliminary version of this work [31] estimated head yaw angle using a limited set of body and head poses involving a motionless subject with a straight body, lowered arms, and a head pose with no roll or pitch rotations. The current work, enhancing the model in [31], is able to estimate orientations while the subject moves naturally and displays various body and head poses. The enhanced model can estimate yaw with similar accuracy even in the presence of roll and pitch rotations. To the best of our knowledge, our head orientation estimation model is the first to estimate orientations from a surveillance viewpoint (low resolution), using LiDAR sensors. Our second contribution in this work is quantifying differences in orienting behavior between neurotypical and autistic individuals using an automated system. Although some of these differences are generally known to characterize ASD, we are the first to quantify them using an automated system, and the first to quantify them in a triadic conversation setting.

The rest of this paper is organized as follows. Section II presents an overview of existing literature on body orientation estimation, head orientation estimation and orientation behavior analysis of autistic individuals. Section III explains our data collection, labeling and cleaning procedures. Section IV presents our methodology starting from data preprocessing and correction, detailing our body orientation estimation procedure as well as our feature extraction and machine learning approach for head orientation estimation. Section V presents the performance of our models in terms of MAE and the comparison with the state-of-the-art. Section V also presents an application of our models where we show significant differences between neurotypical and autistic individuals in terms of orientation and attention distribution behavior. Section VI concludes with a discussion of the current work and our future directions.

II. RELATED WORK

Body and head orientation estimation are well studied tasks in computer vision. In this section, we categorize related work according to the data types (RGB images or depth maps/point clouds) as well as the experiment setting (frontal or surveillance). We then introduce a few studies that analyze and compare orienting behaviors of neurotypical and autistic individuals in similar experimental settings.

A. BODY ORIENTATION ESTIMATION

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 [33] and [34] proposed template matching models that combine 2D images from multiple surveillance viewpoints to make 3D orientation estimates. Studies targeting pedestrian orientation [35], [36] usually approach the problem as a classification task, providing less precision compared to regression models. Many studies such as [37], [38] incorporated motion information and tracking techniques into their models as they approach the task from the perspective of a vehicle. The authors of [4], [39] proposed models to estimate body orientation for human-robot interactions, which resemble a frontal setting.

The authors of [18] proposed a person-tracking system with a body orientation estimation feature using depth sensors from surveillance viewpoints. The authors use Principal Component Analysis (PCA) on projected point clouds to estimate body orientation. Other than [18], the works on body orientation estimation using depth sensors do not use surveillance scenarios. Shimizu *et al.* [40] proposed a model which combines shape and motion information using a LiDAR-mounted robot. Similarly, [41] combines Histogram of Oriented Gradient (HOG) features with motion information tracked by a Kalman filter, using depth images from a Kinect sensor. Other studies use depth along with color information; [42] enhanced features extracted from an RGB image using depth and motion information, while [43] combined features extracted from both RGB and depth images. Experimenting with different CNN architectures that use RGB input, depth input and RGB-D input, authors of [2] argued that depth maps are more suitable for estimating orientation than RGB images.

B. HEAD ORIENTATION ESTIMATION

Many papers estimate all three Euler angles (yaw, pitch, roll) to define a full head pose [5], [6]. Depending on the application, such as behavior prediction on pedestrians [3], [36], pitch and roll angles are often neglected as yaw angle defines the direction people are looking. For our application on the division of attention between two other people in a triadic conversation, we likewise focus on yaw angle.

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. Before deep learning techniques, good results were achieved by [44], [45] using graph embedding, manifold learning and locally linear embedding techniques. Zhao *et al.* [46] used a neural network followed by more complex architectures such as random regression forests [47], deep neural networks [48], [49], convolutional neural networks (CNN) [50]–[52] and Graph-CNNs [53]. The authors of [54], [55] proposed a technique called web-shaped model to estimate head orientation using 68 facial landmarks which are detected using [56]. A recent study by Yao *et al.* [57] showed that state-of-the-art performance can be achieved by using only seven of those 68 landmarks, four of which are the corners of the eyes. All these papers target a frontal setting. The authors of [3], [58] created various models and surveillance settings for the head

orientation estimation task using RGB cameras, as the latter took advantage of extensive research on face detection in a room surveyed by four cameras creating multi-view representations. The authors of [1] proposed RGB image based head orientation estimation models, to ensure pedestrian safety from the perspective of a vehicle. Similarly, a CNN-based model in [36] estimated pedestrian orientations from a surveillance viewpoint. While proposing a transfer learning approach, the authors of [59] published the DPOSE dataset, a dynamic, multi-view head pose dataset collected in a room with 4 cameras at surveillance viewpoints. Various papers [60], [61] proposed new methods and published results on the DPOSE dataset. A CNN-based model in [62] works on unconstrained RGB images, similar to a surveillance viewpoint with higher resolution.

Head orientation estimation using depth cameras has a longer history (see [63]) compared to body orientation estimation using depth cameras. Following the availability of consumer-level depth cameras such as Microsoft Kinect and Intel RealSense, various head orientation estimation models using depth images [5], [6], [15], [16], [64]–[66] were proposed. The BIWI benchmark [5] contains around 15000 samples of head poses recorded with a sensor placed frontally about 1 meter from the subjects, and it was used for many results [6], [15], [16], [53], [60], [67]. A particle swarm optimization approach was used in [16], while [15] used triangular surface patches as hand-crafted 3D features to estimate orientation. More recent papers such as [6], [67] proposed CNN architectures to tackle the problem; [67] resembles our work as their model uses 3D point clouds as input as opposed to the more common 2D depth maps. Point clouds were also used in [68] which leveraged the PointNet++ architecture [69] by using its abstraction layers as a feature extractor, and they further improved their work by including temporal information using an LSTM network in [70]. Similar to the depth information based body orientation estimation models, the models listed above assume that the depth sensor is directly in front of the person.

To our knowledge, we are the first to estimate head orientations using a depth sensor from a surveillance viewpoint. Existing models use either RGB images with a surveillance setting, or use depth images with a frontal setting. In this study, we propose models for both body and head orientation estimation.

C. ANALYSIS OF ORIENTATION BEHAVIOR

A core diagnostic feature of ASD is differences in social attention [71], which include social orienting, joint attention, eye contact, and non-verbal gestures. In this section of our literature review, we mainly focus on orientation behavior, specifically in triadic settings.

As suggested by [72], early triadic behaviors are important for the development of later social responsiveness. The authors of [28] studied triadic conversations with low communicative intent (researchers speaking primarily with each other, with occasional input from a child) and dyadic conversations with high communicative intent (a researcher directly interacting with a child) and found that children with ASD made 57% more gaze fixations to people's faces in these triadic conversations compared to the dyadic ones; the reverse pattern was found for typically developing (TD) children. The authors also found that children with ASD spent 12.3% less time looking at other people's faces in these triadic conversations compared to the dyadic ones, and 9.7% less compared to TD children.

Other studies such as [73], [74] with different experimental settings also provide insight into orienting behaviors of people with ASD. The authors of [73] found that children with autism were significantly less likely to respond to social stimuli (such as calling the child's name, or snapping fingers) with a re-orientation of the head, compared to their responses to non-social stimuli (such as a phone ringing), as well as compared to the responses of TD children. In a virtual public speaking experiment, the authors of [74] found that high-functioning children with ASD made contact with the listeners less frequently compared to TD children.

A model to analyze head movement features such as rotation range and frequency in autistic children during face-to-face interactions was proposed in [75]. The authors reported that compared to TD children, autistic children had a significantly higher level of head movement stereotypy (repetitive, ritualistic head movements), as well as higher rotation range and frequency. Based on these results, the authors developed a machine learning model to diagnose autism in children using the proposed head movement features [76].

Many researchers have studied social modulation of gaze, which is the change in gaze orientation based on conversational role (e.g., speaker or listener). In dyadic conversations, listeners generally gaze more at speakers compared to speakers looking at listeners [77], [78]. However, [78] found that in group conversations, the gaze levels of speakers come close to that of listeners. The authors argued that one reason for this change was that speakers, when addressing a group, need to collect visual feedback from each individual and to maintain the signal that they are addressing each individual.

III. DATA SETS

Our system uses two ToFv2 LiDAR sensors from Hitachi Vantara [79]. The sensors capture depth information and create a point cloud based on the Time-of-Flight principle [80]. We placed sensors at opposite ceiling corners in a 3x3.5 meter conference room, looking down on an oval table. The point clouds are stitched together using rotation and translation. For both the static and conversation datasets, we ensured with calibration tests that the sensor positions, orientation angles and stitching parameters are the same before each data collection session for data reliability. For the static dataset, the sensors were manually calibrated by visual inspection of the output point clouds. We refined the calibration procedure for the conversation dataset using the fixed locations of four pieces of reflective tape that provide stronger sensor signal. We use the sensor software's built-in

human detector which outputs XY-coordinates of the center of gravity of the detected human, as well as a Z-coordinate of the top of the head, Z_{head}. In this section, we present our data collection, labeling and cleaning procedures. This study was approved by the UC San Diego Institutional Review Board (Protocol 210775, Date 7/1/2021).

A. STATIC DATASET

In [31], we created a static dataset from 15 neurotypical adults with and without glasses and face masks, and with varying hairstyles and heights. Our dataset consisted of 8 male and 7 female subjects with an average age of 26.2. We collected data with one subject at a time, while the subject follows guidance arrows (as ground truth) placed on a table. Each subject orients their head towards 13 predetermined angles (-90 to +90 degrees, in increments of 15 degrees). For capturing each point cloud, the subjects are instructed to be motionless with their upper body straight to the front and parallel to the table edge, and with their hands on the table or on their lap. Point clouds corresponding to each head orientation angle were captured one by one as snapshots, rather than getting sampled from a continuous data stream.

B. CONVERSATION DATASET

While the static dataset was useful in the early stages of model development, we found that a model trained on it struggled to accurately estimate head and body orientations of people engaged in real-world conversations, which involve natural movement and varying head and body poses. To create a dataset that represents natural aspects of a conversation, we recorded conversations in a triangular conversational setting with two interviewers and one subject. The subjects were 12 neurodivergent individuals who had received a community diagnosis of ASD and 8 neurotypical individuals. The 12 autistic subjects consisted of 10 males and 2 females with an average age of 22.1 while the 8 neurotypical subjects consisted of 6 males and 2 females with an average age of 23.6. Each subject participated in 2 sessions of 8 to 15 minutes in two different seating setups as shown in Fig. 2. In *Setup*90, the interviewers are separated by an angle of around 90 degrees (ranging from 75 to 105 across sessions) from the subject's perspective, whereas in *Setup*45, the separation is around 45 degrees (ranging between 35 and 55 degrees). Different seating positions helped us collect data with different head orientations, creating useful variety in the dataset.

The subjects were asked to engage naturally as they would in a casual conversation, and they were not informed prior to the session that the data would be used to analyze head and body orientation. The conversation starts with a casual question such as "What do you do in your free time?" and continues based on the answers of the subject. It is intended that the subject is the main speaker throughout the conversation, while the interviewers listen in an engaged way, while also making brief comments, asking follow-up questions, and shifting topics. The duration, pace, topics, and conversational roles were controlled to the best extent possible to prevent these external factors from confounding the behavior analysis portion of the study. The interviewers followed a conversation script reasonably closely so that each participant was asked questions about the same set of topics, in the same order. During the sessions, LiDAR point cloud data were recorded with an average frame rate of 1.5 fps. We also recorded RGB video solely for ground truth labeling and not for model development. Unlike the static dataset, in the conversational setting we observed many different body poses by the subjects, such as turning their upper body towards one interviewer, using their arms and hands as part of their body language, and putting their hands close to their face while thinking or listening.

C. DATA EXTRACTION AND LABELING

To create a dataset that contains various head poses that occur during a real conversation, we manually sampled and labeled data from the sessions. To manually estimate the ground truth head orientation from video snapshots, we used 3 reference orientations. Two of these are computed using the point cloud centroid coordinates of the two interviewers with respect to the subject, at the time of the snapshot. The third reference angle is the average of the first two, representing the midpoint of the two interviewers. For example, if one interviewer is seated 30 degrees to the left of the subject, and the other interviewer is seated 50 degrees to the right, the three reference angles are +30, -50 and -10 degrees. If the subject's head is oriented directly towards an interviewer, the ground truth label is the reference angle for that interviewer. If the subject is looking at the midpoint of the two interviewers (a common situation when speaking to multiple listeners), the ground truth label is the midpoint reference angle. The manual sampling and labeling procedure is detailed below:

1) Align the RGB video and point cloud recordings based on timestamps.
2) From the video recording, identify an instance where the subject's head orientation is static for at least two seconds and close to a reference point.
3) Extract the point cloud data that corresponds to the identified video instance.
4) Calculate the reference angles using the point cloud coordinates of the subject and the interviewers.
5) Estimate the subject's head orientation from the video as ground truth, with the help of reference angles.

From this, we obtained 80 to 140 instances from each ASD subject, totaling 1400 point cloud frames. We ensured a variety of body and head poses in the dataset, including challenging ones such as subjects with their hands on their face or chin, arms over their head or their bodies heavily leaning towards the table, the back or the sides. We also tried to ensure that the frequencies of various different poses within the dataset are reasonably close to how often each pose is displayed by the subjects. We achieved this by sampling a data point every time a subject shifts their pose (e.g. turns

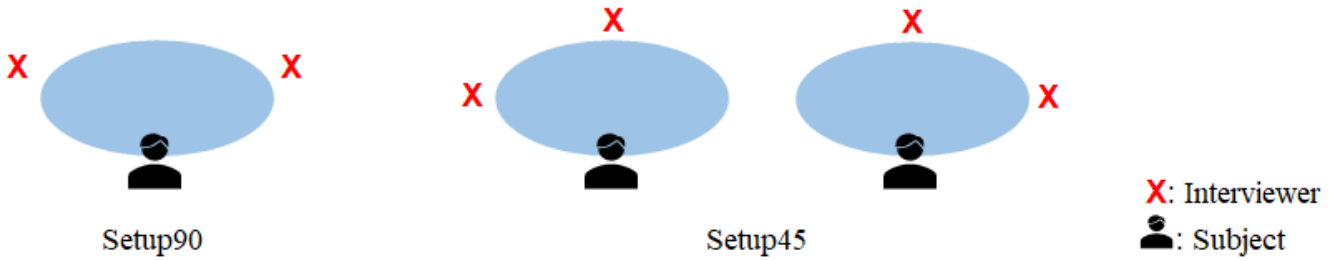

FIGURE 2. Conversation Setups

from one interviewer to the other, leans towards the table, raises their arm, places their hands on their face etc.), and sampling data points periodically (one sample every 10 to 15 seconds) if the subject's pose is stable for longer periods. Note that sampling a data point is only possible if there is an instance where the subject's head orientation is static for at least two seconds and close to a reference point, as stated by the second item in the above data sampling procedure. It is also important to note that this process is done only once to establish ground truth for creating the model, and should not be construed as a calibration step needed in subsequent use of the model.

This manual sampling and labeling is subject to potential human errors. Estimating head orientation accurately from a video is hard, although we sampled instances where the subject's head is oriented towards an interviewer or the midpoint, which are relatively easier positions to interpret the orientation. To quantify the human labeling error, we conducted a simple experiment on three subjects. Using guidance arrows to guide the subject to adjust their head orientation in 5 degree increments, we collected a random sequence of head orientations. On average, human labels based on video recordings differed by about 4.5 degrees from the guidance arrow ground truth. A portion of this error comes from the subjects imperfectly aligning their heads with the guidance arrows, which is a potential issue that exists in the static dataset as well.

D. DATA CLEANING

Our dataset presented challenges in data cleaning. First, there were issues caused by the sensors. Network overload during real-time data collection caused lost point cloud frames that were replaced by previously recorded frames, an issue present in 3 of 24 data collection sessions. Of the 1400 point clouds manually sampled from the real-time sessions, 56 were excluded as they were sampled from a repetition sequence and therefore did not reflect the true state of the environment at the matched timestamp. Another LiDAR sensor issue was the inaccuracy of the built-in human detection algorithm. The sensor may confuse the subject's shoulder with their head, resulting in a wrong output of center of gravity and Z_{head}. In that case, the wrong portion of the human body is cropped out and the point cloud lacks points from the other shoulder. We removed an additional 74 point clouds due to this erroneous human detection. In Section IV-B, our proposed head position estimation algorithm can mitigate this issue. We also used our proposed algorithm to detect the instances where this issue happened and eliminate them if the discrepancy between the built-in head position and our estimated head position is bigger than 10 cm either in the horizontal plane or in the vertical axis.

Secondly, there are human errors during manual sampling and labeling, such as selecting a wrong timestamp from the video recording, or a slight lack of synchronization between the video recording and point cloud sequence, leading to the selection of the wrong point cloud frame. In such instances, we observed that the next or previous point cloud is more suitable for the suggested head orientation label, indicating that the wrong point cloud was sampled and the subject's head orientation changed within consecutive frames. We address this problem by using our model's predictions on the neighboring frames as a preliminary indication of a wrong sampling or a synchronization issue, and replace the point cloud with its neighbor if we can visually confirm the issue from the video recording and point cloud sequence. In Fig. 3, two consecutive point clouds from one of our data sequences are shown. The human labeler originally sampled Fig. 3a from the sequence, trying to match with the video instance where the subject's head was towards an interviewer seated 35 degrees to the subject's left. However, due to a slight synchronization issue, the point cloud in Figure 3a belongs to the middle of the head movement towards that interviewer, and the next point cloud (shown in Figure 3b) should have been sampled instead. The head movement becomes apparent when 4-5 consecutive point clouds are visualized on top of each other and compared with the corresponding video sequence, which allows one to choose the point cloud matching the intended orientation label. Of the remaining 1270 point clouds, 41 were replaced by their neighboring point clouds due to this labeling issue.

IV. METHODOLOGY

In this section, we present our data pre-processing steps and body and head orientation estimation models. Section IV-A details the modified noise removal algorithm from [31] and Section IV-B presents a head position correction procedure which led to improvements in model performance. Section

IV-C presents our body orientation algorithm. Section IV-D describes our hand-crafted geometric features and feature elimination procedure while Section IV-E presents our head orientation estimation pipeline.

A. PRE-PROCESSING

For this work, we introduced additional pre-processing steps compared to our work in [31]. As explained in Section IV-B, the built-in head height estimation is often inaccurate, so we propose an improved estimation procedure to obtain the head height, Z_{head}. To extract the region of interest which consists of the upper body and head, we crop a cylinder-shaped boundary around each person's point cloud using the centroid and a radius of 50 cm. For each subject, a threshold for the upper body set empirically as the top 27% of their height (in a seated position) is computed from our improved estimated Z_{head}. We also removed points from the head point cloud if they are at least 15 cm away from the head center and from the upper body point cloud if they are at least 25 cm away from the body center on the horizontal plane, after separating the head and upper body point clouds. The computation of head and body centroids and the separation of head and upper body point clouds are explained in Section IV. With these additional steps, we were able to remove points that did not belong to the region of interest and instead belong to the table, the back of the chair, or noise, which created many distortions in our preliminary work [31].

The upper body point clouds obtained from the pre-processing step consist of about 1800 points on average, varying between about 1500 and 2100 points per case. Since our system is in a surveillance setting, our upper body point clouds have lower resolution compared to other work, e.g., the BIWI dataset [5] contains around 10,000 points for a person's face alone. Estimating body and head orientation from low-resolution LiDAR data is challenging due to the lack of detail in the small region of interest. Moreover, the point cloud data from the surveillance angle are noisy, especially from hair and other complex features on the head. To mitigate this, we apply a k-nearest neighbor noise removal step, where we delete a point if the average distance between the point and its 10 nearest neighbors is larger than 50 mm. All the parameters and thresholds presented in this section are treated as hyperparameters which were optimized during the training of our head orientation estimation model. We initialized each parameter based on our visual and statistical analysis of the data, and optimized them for model performance.

B. HEAD POSITION CORRECTION

Although the LiDAR's built-in human detection capability usefully extracts human point clouds from the environment point cloud, it does not pinpoint the head center in the horizontal plane as the center of gravity of the human is not necessarily the same as their head center. The algorithm also provides inconsistent results for Z_{head}. The two sensors make independent estimates which are averaged to form a joint estimate, which is usually better than relying on a single sensor estimate. However if one sensor makes a large estimation error, the joint estimation is not good enough to recover. Accurate and consistent estimation of Z_{head} across the whole dataset is especially important as it is used to separate the head and upper body point clouds.

We improved the estimate of the head center and Z_{head} from the point cloud. If Z_1 and Z_2 (in centimeters) represent the built-in Z_{head} estimates for sensors 1 and 2, we use $Z = ((Z_1 + Z_2)/2) \quad 15$ as the initial separation threshold; points above this threshold belong to the head and points below belong to the upper body. This yields two disjoint point clouds, PC_{head} and PC_{body}. We project the points in PC_{head} onto the horizontal plane and use least-squares ellipse fitting on them, as detailed in the following section. The ellipse center is a more accurate estimate of the head center, compared to the built-in estimate from the LiDAR sensors. Z_{head} is determined by sorting the points by their z-coordinate and taking the highest point with a maximum height difference of 1mm with the next 5 highest points. This operation mitigates noise distorting the Z_{head} calculation, and generally pinpoints the top of the head where the height difference between points should be saturated. The head center computed from the least-squares ellipse, together with this Z_{head}, represent the center point of the subject's top of the head.

This improves our preliminary work [31], which relied heavily on the built-in estimates. In [31], the inaccurate built-in estimation for Z_{head} was used to base the separation threshold to obtain PC_{head} and PC_{body}, which sometimes caused PC_{head} to contain points from the shoulders or PC_{body} to contain points from the chin.

C. BODY ORIENTATION ESTIMATION

The body orientation estimation model is a geometric model which takes advantage of the ability to change the viewpoint from which a point cloud is seen, and uses the birds-eye view of the room. The cropped point clouds are projected onto the horizontal plane. After estimating Z_{head}, we separate head and body points using a refined threshold of $Z = Z_{head} \quad 17{:}5$ and calculate the 2D ellipse that best fits the projected PC_{body} based on least squares error, with the long axis of the ellipse representing the frontal (shoulder-to-shoulder) axis. We use the conic representation of an ellipse:

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

The optimal coefficients are estimated using the direct least squares ellipse fitting method by Fitzgibbon *et al.* [81]. The noise removal pre-processing is important for this procedure to work well, as noise points that are generally on the edges may result in large squared errors. The correction of the built-in Z_{head} estimation is also crucial as explained in the previous section.

After the frontal axis is determined, there remains the issue of which side of the ellipse is the front. We calculate the average perpendicular distance of each point in PC_{head} from

both sides to the frontal axis. Assuming that a person's head is almost always in front of their body (their frontal axis), the front is taken as the side with higher average perpendicular distance. In our datasets, this assumption holds true 99.7% of the time and the front side is correctly determined.

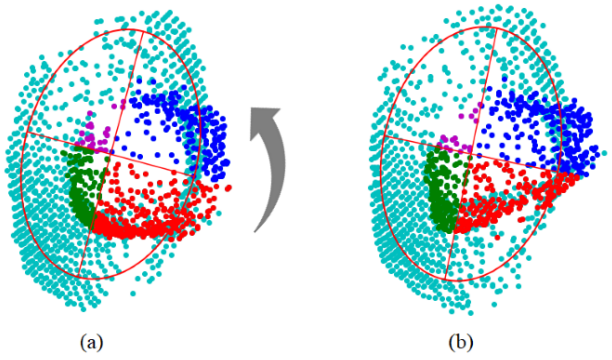

FIGURE 3. Least squares ellipse fitting for body orientation estimation via the long axis of the fitted ellipse, which also determines the four quadrants of the head relative to the body. Light blue points are projected upper body points (shoulders and chest); dark blue, red, purple and green points are projected head points, representing the four quadrants, in order. The first and second quadrants represent the front-left and front-right sides of the head, respectively; while the third and fourth quadrants represent the back-left and back-right sides of the head. (a) Head orientation label unknown; point cloud belongs to a head movement that starts at 0 degrees and moves to the left (b) Head orientation labeled as 35 degrees to the left.

D. FEATURE EXTRACTION AND SELECTION

For head orientation estimation, simple geometric approaches were not sufficient as details of facial features are not accurately captured by the sensors. Therefore, we engineered our own geometric features from the point clouds. The surveillance setting allows low-level geometric features but, given the low-resolution nature of our data, higher level 3D features such as surface patches [5], [15] or curvatures [82] did not produce useful results on our dataset. The approach proposed in [68], using the initial layers of PointNet++ to extract feature representations, or a Graph-CNN approach proposed in [67] similarly did not suit our data. However, as shown in our prior work [31] and the current study, our low-level geometric features allow estimation of head orientation with reasonable accuracy in a low-resolution surveillance setting.

The feature extraction is done after noise removal and ellipse fitting to the upper body. The upper body ellipse divides PC_{head} into four quadrants which produce supportive features for the model, based on the point locations with respect to the body center. Fig. 3 shows two projected human point clouds, the optimal ellipse fit for body orientation estimation, and the resulting four quadrants of the head for each of the clouds.

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 points in PC_{head} (mean, standard deviation, minimum and maximum coordinates), as well as of the points in its four quadrants separately, 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 PC_{head}. Some of the features we extracted emerged from our initial ideas on how to achieve accurate head orientation estimations. For example, the principal components of the head point cloud and the furthest points from the head centroid corresponding to the nose tip were two ideas to directly estimate head orientations. While none of these worked well on their own, they served well as features to a more complicated model. The features extracted from a single point cloud constitute a feature vector with 103 entries (as x and y dimensions produce distinct features).

For a low-resolution regression task, a feature space with over 100 dimensions presents a higher likelihood of overfitting. To mitigate this, we use the Random Forest Recursive Feature Elimination (RF-RFE) process [83] which involves repeatedly training a random forest regressor, ranking the features according to their importance, and eliminating the least important feature(s) in each iteration. This approach has been successfully used in many studies [84]–[86]. After applying RF-RFE , the optimal feature set had 42-dimensions, with principal components proving to be important features along with some engineered features such as the estimated nose position and the head ellipse parameters. The feature elimination procedure revealed that our two initial ideas involving principal components and nose tip estimation were among the most useful features. Other features in the optimal feature set are a mix of head quadrant principal components and distribution properties of the head point cloud in certain dimensions. Intuitively, horizontal dimensions should hold more importance compared to the vertical dimension since we are estimating the head orientation in the yaw axis. Some head quadrants turned out to be more important than others based on where the sensors are located and their angles in which they view the subjects.

E. HEAD ORIENTATION ESTIMATION

For head orientation estimation, we use a pipeline of feature extraction and an ensemble of multi-layer perceptron-based regression networks. To train the head orientation estimation model, we use leave-one-out cross-validation, where the point clouds of each subject are used one time as the test set, and used in training otherwise. Thus each autistic subject has their own model that has never seen that subject before. Depending on the subject, each of the leave-one-out models was trained using 1150 to 1200 point clouds from the dataset and tested on 70 to 120 point clouds. On average, each model was trained with 92% of our dataset. For neurotypical subjects, we use a model trained with the whole dataset of samples from autistic subjects.

We chose to use a neural network based regression model as a result of experimentation with multiple different ap-

TABLE 1. Conversation Dataset MAE for Head Orientation Estimation after each Data Cleaning and Processing Step

Process	Mean Absolute Error on Conversation Dataset	
	Model Trained with Conversation Dataset	Model Trained with Static Dataset
Initial Model (from [31] but trained as specified in column)	19.6	35.1
Head position correction (Sec. IV-A)	18.3	32.8
Removing repeated point clouds (Sec. III-D)	17.2	32.8
Removing point clouds with missing points (Sec. III-D)	15.9	29.5
Using neural network ensembles (Sec. IV-E)	14.2	26.4
Fixing wrongly synchronized point clouds (Sec. III-D)	13.7	26.4

proaches and algorithms. As discussed in Section IV-D, we initially experimented with other proposed head orientation estimation algorithms from the literature. After concluding that the existing approaches are not suitable for our dataset, we extracted our own features and experimented with multiple different machine learning algorithms such as SVM, decision trees, random forests, gradient boosting and neural network based regression. The latter performed best, with random forest regression a close second. We used random forest regression to measure feature importance as discussed in Section IV-D.

Neural networks are typically high variance estimators, as was our preliminary model [31]. A dataset of noisy low-resolution point clouds leads to even more variance in predictions. To reduce the estimation variance and improve overall model performance [87], [88], we enhanced our initial model by deploying an ensemble of neural networks, where each individual network was initialized with different random weights. Often, different initial weights are enough to generate significantly different models [87], [88], to create a diverse ensemble. To create the ensemble, we train 20 separate models and rank them based on their performance on the validation set. Then we use Forward Subset Selection [87] to select the models as follows. We start with an initial ensemble of 3 best models, and iteratively add the next best model in the pool to the ensemble until the ensemble performance on the validation set does not improve with the addition of a new model. We ended up with ensembles that contain 3 to 8 models, with the mode and median being 6 models.

V. RESULTS

In this section, we discuss the performance of our proposed models. Section V-A evaluates our models based on mean absolute error (MAE). We present MAE values for our models based on the number of features, the selected feature set and an ablation study of each of the pre-processing steps. We also compare our work to two state-of-the-art head orientation estimation models, as well as other existing literature. Section V-C introduces an application of our estimation models, comparing attention distribution patterns of neurotypical and autistic individuals in triadic conversation settings. We find statistically significant differences between the two groups.

A. ERROR METRICS

We primarily use MAE to evaluate model performance. For the body orientation estimation model applied to the static dataset, with our proposed improvements, we achieve an MAE of 5.21 degrees compared to the MAE of 8.37 degrees reported in [31]. The model improved significantly with the head position correction presented in Section IV-B. Our proposed ellipse fitting method for body orientation estimation outperforms the PCA approach proposed by [18], as the latter produced an MAE of 7.95 on our dataset. We found that the ellipse fitting approach is more robust against noise in the point clouds.

We train and evaluate the head orientation estimation model on data sampled from our new conversation dataset, from 12 autistic subjects. We use a different model for each subject, where data from the other 11 subjects are used for training. On average, our new modeling approach produces an MAE of 13.73 degrees across 12 leave-one-out models for head orientation estimation. When the model from [31] that was trained with only the static dataset is applied to the 12 subjects on the conversation dataset, the MAE is 26.4 degrees due to the challenges caused by different body poses and natural movements in real-world settings. Our new model based on our conversation dataset outperforms our model in [31] by about 50% in a conversational setting. In [31], we reported an MAE of 12.69 on our experimental static setting, showing that our new model is able to reach similar levels of accuracy in a conversational setting. A more detailed summary is presented in Table 1 which shows the development of our final model as well as a comparison with our initial model.

Fig. 4 shows the evolution of model performance (MAE) as we eliminate features with the RF-RFE procedure described in Section IV-D. Without RF-RFE, using the whole feature space, the model performance would have been 13% worse compared to the optimal feature set, with an MAE of 15.72 degrees. The MAE of the model with only 1 feature is 25.61 degrees, 46% worse than the optimal performance.

After obtaining the optimal feature set using the RF-RFE procedure, we conducted some experiments to analyze feature importance. Fig. 5 compares model performance in terms of MAE with the optimal feature set and its subsets in which certain features are excluded. The figure shows the importance of engineered features, including the estimated nose position, principal components of the head quadrants, and parameters of the head ellipse.

TABLE 2. State-of-the-Art Comparisons for Head Orientation Estimation

Model	MAE on our dataset	MAE on respective dataset
Proposed Neural Network Ensemble	13.7	-
PointNet++ Regression [68]	24.2	7.32 (MSE)
Graph-CNN PointNet++ [67]	23.5	1.82
HOG-SVM-HMM Pipeline [1]	-	19
Coupled Adaptive Classifier [3]	-	23.6
Multitask Manifold CNN [60]	-	31
Face Detection-Naive Bayes-HMM [58]	-	33.6

TABLE 3. Comparison of neurotypical and autistic subjects on two experimental setups

Statistic	Setup90		Setup45	
	ASD	NT	ASD	NT
Average duration of contact	11.8	6.3	7.7	6.1
Maximum duration of contact	48.8	16.5	24.6	16.5
Average duration of NOT contacting anyone	53.2	41.1	58.6	30.9
Number of contacts per minute	0.84	1.37	1.44	2.15
Total duration of contact % during an interview	11.5%	13.4%	15.6%	16.6%
Maximum duration of exclusions	50.3	16.4	20.3	9.3
Number of exclusions per minute	0.28	0.10	0.19	0.11
Total duration of exclusions % during an interview	15.5%	4.2%	7.4%	3.1%

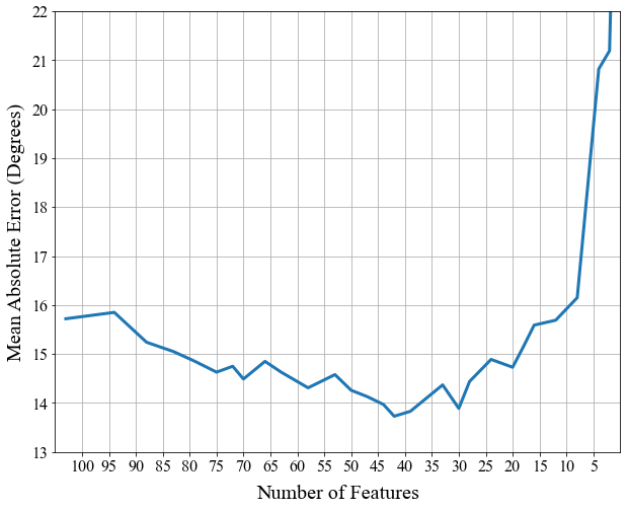

FIGURE 4. Evolution of mean absolute estimation error with the Random Forest Recursive Feature Elimination procedure. The initial MAE with 103 features is 15.72 degrees, whereas the MAE with only 1 feature left in the feature space is 25.61 degrees. The optimal feature set contains 42 features and leads to an MAE of 13.73 degrees.

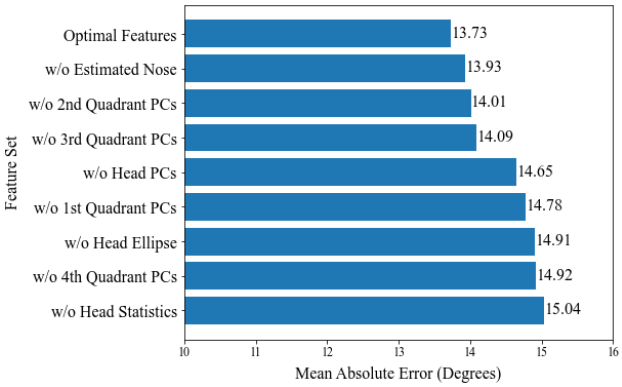

FIGURE 5. Performance of optimal feature set and its subsets (PC = Principal Component).

B. COMPARISONS WITH STATE-OF-THE-ART

While some papers report smaller errors on head orientation estimation, they use either high-resolution 3D scans of the face when the sensor is placed directly in front of the person [5], [6], [15], [16], [64]–[66], or an RGB camera [47]–[50]. But for those studies which used indoor RGB surveillance datasets, MAE values of 19, 23.6, 31 and 33.6 degrees were reported [1], [3], [60] and [58], respectively. While we outperform the above studies which used RGB cameras from surveillance viewpoints in terms of raw MAE numbers, it is hard to make exact comparisons as the datasets are not unified and each dataset has its own challenges.

To make fair comparisons, we evaluate the performance of two PointNet++ based state-of-the-art architectures on our dataset. The architecture in [68] uses the set abstraction layers of PointNet++ as feature descriptors followed by a fully connected regression layer. The authors of [67] modified the PointNet++ set abstraction layers with a Graph-CNN approach, proposing a cascaded classification and regression architecture. Both of these architectures rely on PointNet++'s ability of extracting local features from a point cloud. Since these local features may not be easily identifiable in low-resolution point clouds, these approaches did not work as well on our dataset, producing MAE values of 24.2 and 23.5 degrees for [68] and [67], respectively.

C. BEHAVIOR ANALYSIS OF AUTISTIC AND NEUROTYPICAL INDIVIDUALS

In this section, we quantify some orienting behaviors of autistic and neurotypical individuals in two experimental triadic conversation setups, as shown in Fig. 2.

To quantify behaviors, we define the following two terms: *Contact* and *Exclusion*. Based on studies that suggest "2 to 3 seconds" [89], [90] or "a few seconds" [91] of eye contact is optimal when addressing multiple people to connect and make them feel included in the conversation, *Contact* in our context is defined as 3 consecutive frames where the head orientation is inside the region of an interviewer, where a frame is about 0.7 seconds and an interviewer's region is defined as 15^o from their position. Making occasional contact with an interviewer is important to make them feel included in the conversation [92] and maintain the signal that they are being addressed [78]. *Exclusion* is defined over a 20-frame window; if there are at least 15 estimated head orientations in the region of one interviewer and none in the region of the other, the other interviewer is considered to be excluded from the conversation. In Table 3, we present statistics related to *Contact* and *Exclusion* extracted from our conversation sessions using our head orientation estimation model. We present the averages for each statistic across 12 sessions with autistic subjects and across 8 sessions with neurotypical subjects in each setup.

From Table 3, we observe that the orienting behaviors of the autistic and neurotypical individuals diverge more in $Setup90$, compared to $Setup45$. When the interviewers are further apart, the autistic individuals have more difficulty with distributing their attention between two conversational partners. Neurotypical subjects tend to make contact with the interviewers in shorter bursts, whereas autistic individuals frequently dwell on one interviewer for a longer period of time. With shorter and more frequent contacts, neurotypical individuals are more likely to ensure that both interviewers feel included in the conversation. Autistic individuals more often have an exclusion, seen through the much higher maximum exclusion duration and total percentage of time spent while an interviewer is excluded.

The authors of [74] suggested that children with ASD made fewer contacts with listeners while speaking to multiple people, compared to typically developing children. Similarly, the authors of [28] showed that autistic children spend less time looking at other people's faces in triadic conversations compared to TD children. Our findings are consistent with these, as we observe from Table 3 that people with ASD made fewer contacts per minute and spent less total time in contact with the listeners.

We examined statistical significance with independent sample t-tests on the data of the two groups. An independent t-test suggests that there is a significant difference between the averages of two groups if the p-value (the probability of this difference occurring by chance) is smaller than the widely accepted threshold of 0.05. We also report Cohen's d values associated with each statistic as the effect size, which is the difference between the group means divided by the pooled standard deviation [93]. Cohen's d values of 0.2, 0.5, 0.8 generally correspond to small, moderate and large differences between two groups, respectively. A summary of behaviors we found to be significantly different between the two groups can be found in Table 4.

In *Setup90*, we found that the maximum duration of exclusions and total duration of exclusions percentage during an interview exhibit significantly different results between the two groups. No significant difference was observed for the *Exclusion* statistics in *Setup45,* supporting the idea that distributing attention was harder in *Setup90* compared to *Setup45* for the ASD participants. Among *Contact* statistics, the average duration of not making contact with anyone was significantly higher in *Setup90* for autistic individuals in comparison with neurotypical individuals. The number of contacts per minute was significantly lower for people with ASD, in both *Setup90* and *Setup45*.

In Tables 5 and 6, we present the distributions of head orientations based on the subject's conversational role. In this analysis, we again observe that the differences between the two groups are more evident in $Setup90$. The table shows that, when speaking to multiple people, autistic individuals tend to distribute their attention less evenly; their focus usually remains on the person who made the last remark. Neurotypical individuals pay slightly more attention on the person who spoke last, while generally ensuring that the other interviewer is also included in the conversation. This difference is confirmed to be significant by t-tests, as the results reveal that people with ASD tend to look at the interviewer who spoke last significantly more than neurotypical people do. To the best of our knowledge, this is the first quantification of these types of differences about conversational roles and their impact on orienting in triadic settings.

Overall, we conclude that there are noticeable differences between the two groups, and our model is able to reflect and quantify these differences. This is valuable towards the goal of creating a coaching tool which would allow autistic individuals to undertake situational practice. While there are many studies regarding social communication behaviors of autistic people, none of them address these in the context of three-way conversations among adults. The extensive literature on dyadic interactions generally found that autistic individuals, compared with neurotypical individuals, display behavioral differences such as spending less time looking at other people's faces [71] and providing fewer nonverbal cues such as regular eye contact or maintaining a body orientation towards a speaker [23]. Based on the literature on dyadic interactions, and the limited literature on triadic interactions for children [28], [74], one can expect that autistic adults would display these different attention distribution behaviors in triadic settings as well. As far as we know, we are the first to conduct experiments that characterize differences between neurotypical and autistic adults in triadic conversational settings, and our findings align with these expectations. Although our small and non-random sample does not allow

TABLE 4. Summary of statistically significant behavioral differences between the ASD and NT populations

Behavior	Setup	ASD Mean	NT Mean	t-statistic	p-value	Cohen's d
Maximum duration of exclusions	Setup90	50.3	16.4	2.53	0.014	1.24
Exclusions % during an interview	Setup90	15.5%	4.2%	2.69	0.011	1.41
Average duration of not making contact with anyone	Setup90	53.2	41.1	1.84	0.046	0.66
Number of contacts per minute	Setup90	0.84	1.37	-2.21	0.023	1.08
Number of contacts per minute	Setup45	1.44	2.15	-1.95	0.037	1.34
While speaking, looking at the interviewer who spoke last	Setup90	30.3%	22.2%	1.87	0.038	0.45

TABLE 5. Distribution of head orientations while subject is listening

Head Orientation - Listening	Setup90		Setup45	
	ASD	NT	ASD	NT
Interviewer who is speaking	27.1%	29.2%	27.4%	28.4%
Neutral	67.6%	67.4%	65.5%	65.6%
Other interviewer	4.9%	3.7%	5.5%	4.9%

TABLE 6. Distribution of head orientations while subject is speaking

Head Orientation - Speaking	Setup90		Setup45	
	ASD	NT	ASD	NT
Interviewer who spoke last	30.3%	22.2%	22.1%	18.9%
Neutral	55.2%	60.8%	63.0%	64.7%
Other interviewer	14.1%	17.1%	13.4%	15.3%

generalization, this preliminary quantification of group differences using an automated system shows an application of our estimation models and is one of our contributions.

VI. CONCLUSION AND FUTURE WORK

In this paper, we improve our proposed models in [31] for body and head orientation estimation that work with low-resolution point clouds generated by two LiDAR sensors. We improve the average error rate of our body orientation estimation model from 8.4 degrees to 5.2 degrees. We enhance our head orientation estimation model by enabling reliable estimations in realistic scenarios where the subject is naturally moving with various head and body poses in a triadic conversation setting. We present novel models that are the first to reliably estimate body and head orientations using LiDAR sensors from surveillance viewpoints. Our estimation results are comparable to results in the literature, although our models work with low-resolution and noisy point clouds and without color information. We also showed that the state-of-the-art models perform poorly in our low-resolution setting although they are effective in high-resolution datasets. This work pushes the boundaries of current body and head orientation estimation systems by demonstrating for the first time that low-resolution, noisy point clouds from LiDAR sensors, without color information, can be used to estimate both body and head orientations from surveillance viewpoints. We believe that accurate orientation estimation models that can work from unobtrusive distances are a significant development. Our proposed models could help with pedestrian safety [1], [17], behavior analysis and prediction [3], interaction and attention modeling [18]–[20], while protecting user privacy.

As an application of our head orientation estimation model, we created a triadic conversation scenario in a room with LiDAR sensors placed to surveillance viewpoints. Using our proposed model, we provide novel analysis on various behaviors in triadic interaction settings and show the differences between autistic and neurotypical individuals using statistical significance tests. We are the first to quantify these qualitatively well-known behavioral differences.

Limitations of this study; In future iterations of this work, we can improve data collection; the network overload issue can be prevented by using a sensor that supports a 1 Gbps network instead of 100 Mbps, as well as a more powerful CPU. The estimated head positions from the built-in human detection algorithm will be corrected using our head detection algorithm presented in Section IV-B.

An additional limitation of this study is the small size of the dataset; our conversation dataset has 12 autistic and 8 neurotypical individuals. While we found some statistically significant differences, other differences might be revealed with a larger number of subjects. We do not claim that our small and non-random sample is representative of the larger population of employment-seeking autistic young adults. Our purpose with this comparison is to highlight the functionality of the proposed technological contribution that is the orientation estimation models, and it also provides preliminary baseline results for future studies.

A potential limitation of this study is scalability and robustness to different environments, which we have not experimented with yet for the current study, but plan to explore in our future studies. Another limitation is the environmental control. While all conversations were held in the same room under the same conditions of heat and light and seating arrangements, the naturalistic flow of conversation led to some conversation variability across subjects. Although the interviewers tried to follow conversational scripts closely, minor differences in conversation flow might have affected subject behavior patterns.

Applications: The proposed body and head orientation estimation models can be used in various applications. We plan to extend our models to become a component of virtual coaching to high-functioning autistic individuals who are seeking jobs, to integrate them to workplaces. We plan to deploy a behavioral intervention program for autistic individ-

uals using our proposed head orientation estimation and behavior analysis tools so that we can further test their effectiveness, while also getting feedback from our participants about this technology. We also plan to develop automated smart coaches that leverage the differences between neurotypical and autistic individuals, while imitating the decisions of a professional behavioral coach. In conjunction with employer-based initiatives to make workplace environments and hiring practices more autism-friendly, tools that allow situational practice and feedback of social communication could facilitate transition to employment for the large number of autistic individuals aging into adulthood each year.

ACKNOWLEDGMENT

We thank Ara Jung, Trent Simmons, Sarah Luo and Saygin Artiran for helping with data collection, Sarah Hacker and Ara Jung for managing the IRB approval and subject payments, and all the participants in our study.

REFERENCES

[1] E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *17th Int. IEEE Conf. Intell. Transp. Syst. (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 *2019 IEEE/RSJ Int. Conf. Intell. Robots and Syst. (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 Conf. Comput. Vis. and Pattern Recognit. (CVPR)*. IEEE, 2012, pp. 1544–1551.

[4] Y. Kohari, J. Miura, and S. Oishi, "CNN-based human body orientation estimation for robotic attendant," in *IAS-15 Workshop on Robot Perception of Humans*, 2018, Art. No. 1.

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

[6] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 42, no. 3, pp. 596–609, 2018.

[7] L. Stark, A. Stanhaus, and D. L. Anthony, ""I don't want someone to watch me while I'm working": Gendered views of facial recognition technology in workplace surveillance," *Journal Assoc. for Inf. Sci. and Technol.*, vol. 71, no. 9, pp. 1074–1088, 2020.

[8] D. P. Bhave, L. H. Teo, and R. S. Dalal, "Privacy at work: A review and a research agenda for a contested terrain," *Journal of Manage.*, vol. 46, no. 1, pp. 127–164, 2020.

[9] J.-E. Joo, Y. Hu, S. Kim, H. Kim, S. Park, J.-H. Kim, Y. Kim, and S.-M. Park, "An indoor-monitoring LiDAR sensor for patients with alzheimer disease residing in long-term care facilities," *Sensors*, vol. 22, no. 20, p. 7934, 2022.

[10] "Lidar market: Forecast and analysis 2022-2028: Skyquest." [Online]. Available: https://skyquestt.com/report/lidar-market

[11] M. Bouazizi, C. Ye, and T. Ohtsuki, "Activity detection using 2D LIDAR for healthcare and monitoring," in *2021 IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2021, pp. 01–06.

[12] A. Günter, S. Böker, M. König, and M. Hoffmann, "Privacy-preserving people detection enabled by solid state LiDAR," in *2020 16th Int. Conf. Intell. Environ. (IE)*. IEEE, 2020, pp. 1–4.

[13] B. Rodrigues, L. Müller, E. J. Scheid, M. F. Franco, C. Killer, and B. Stiller, "LaFlector: a privacy-preserving LiDAR-based approach for accurate indoor tracking," in *2021 IEEE 46th Conf. Local Comput. Netw. (LCN)*. IEEE, 2021, pp. 367–370.

[14] C. N. Lokugam Hewage, D. F. Laefer, A.-V. Vo, N.-A. Le-Khac, and M. Bertolotto, "Scalability and performance of LiDAR point cloud data management systems: A state-of-the-art review," *Remote Sens.*, vol. 14, no. 20, p. 5277, 2022.

[15] 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 *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2015, pp. 4722–4730.

[16] P. Padeleris, X. Zabulis, and A. A. Argyros, "Head pose estimation on depth data based on particle swarm optimization," in *2012 IEEE Comput. Soc. Conf. Comput. Vis. and Pattern Recognit. Workshops*. IEEE, 2012, pp. 42–49.

[17] W. Wang, X. Chang, J. Yang, and G. Xu, "Lidar-based dense pedestrian detection and tracking," *Appl. Sci.*, vol. 12, no. 4, p. 1799, 2022.

[18] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *IEEE Tran. Human-Mach. Syst.*, vol. 43, no. 6, pp. 522–534, 2013.

[19] X. Liu, N. Krahnstoever, T. Yu, and P. Tu, "What are customers looking at?" in *2007 IEEE Conf. Adv. Video and Signal Based Surveillance*. IEEE, 2007, pp. 405–410.

[20] C.-W. Chen, R. C. Ugarte, C. Wu, and H. Aghajan, "Discovering social interactions in real work environments," in *2011 IEEE Int. Conf. Autom. Face & Gesture Recognit. (FG)*. IEEE, 2011, pp. 933–938.

[21] S. O. Ba and J.-M. Odobez, "A study on visual focus of attention recognition from head pose in a meeting room," in *Int. Workshop Mach. Learning for Multimodal Interaction*. Springer, 2006, pp. 75–87.

[22] R. Stiefelhagen, J. Yang, and A. Waibel, "Estimating focus of attention based on gaze and sound," in *Proc. 2001 Workshop on Perceptive User Interfaces*, 2001, pp. 1–9.

[23] 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.

[24] J. L. Chen, G. Leader, C. Sung, and M. Leahy, "Trends in employment for individuals with autism spectrum disorder: A review of the research literature," *Review Journal of Autism and Develop. Disorders*, vol. 2, no. 2, pp. 115–127, 2015.

[25] E. Müller, A. Schuler, B. A. Burton, and G. B. Yates, "Meeting the vocational support needs of individuals with asperger syndrome and other autism spectrum disabilities," *Journal of Vocational Rehabil.*, vol. 18, no. 3, pp. 163–175, 2003.

[26] M. G. Greene and R. D. Adelman, "Beyond the dyad: communication in triadic (and more) medical encounters," *The Oxford Handbook of Health Commun., Behav. Change, and Treatment Adherence*, pp. 136–154, 2013.

[27] A. Nagels, T. Kircher, M. Steines, and B. Straube, "Feeling addressed! the role of body orientation and co-speech gesture in social communication," *Human Brain Mapping*, vol. 36, no. 5, pp. 1925–1936, 2015.

[28] A. McParland, S. Gallagher, and M. Keenan, "Investigating gaze behaviour of children diagnosed with autism spectrum disorders in a classroom setting," *Journal of Autism and Develop. Disorders*, vol. 51, no. 12, pp. 4663–4678, 2021.

[29] A. P. Association et al., *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Assoc. Washington, DC, 2013, vol. 5.

[30] W. H. Organization et al., *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, 1992.

[31] O. N. Tepencelik, W. Wei, L. Chukoskie, P. C. Cosman, and S. Dey, "Body and head orientation estimation with privacy preserving LiDAR sensors," in *2021 29th Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 766–770.

[32] S. Artiran, L. Chukoskie, A. Jung, I. Miller, and P. Cosman, "HMM-based detection of head nods to evaluate conversational engagement from head motion data," in *2021 29th Eur. Signal Process. Conf. (EUSIPCO)*. IEEE, 2021, pp. 1301–1305.

[33] L. Chen, G. Panin, and A. Knoll, "Human body orientation estimation in multiview scenarios," in *Int. Symp. Vis. Comput.* Springer, 2012, pp. 499–508.

[34] M. C. Liem and D. M. Gavrila, "Person appearance modeling and orientation estimation using spherical harmonics," in *2013 10th IEEE Int. Conf. and Workshops Autom. Face and Gesture Recognit. (FG)*. IEEE, 2013, pp. 1–6, doi: 10.1109/FG.2013.6553728.

[35] H. Liu and L. Ma, "Online person orientation estimation based on classifier update," in *2015 IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2015, pp. 1568–1572.

[36] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, 2018.

[37] F. Flohr, M. Dumitru-Guzu, J. F. Kooij, and D. M. Gavrila, "A probabilistic framework for joint pedestrian head and body orientation estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1872–1882, 2015.

[38] I. Ardiyanto and J. Miura, "Partial least squares-based human upper body orientation estimation with combined detection and tracking," *Image and Vis. Comput.*, vol. 32, no. 11, pp. 904–915, 2014.

[39] C. Weinrich, C. Vollmer, and H.-M. Gross, "Estimation of human upper body orientation for mobile robotics using an SVM decision tree on monocular images," in *2012 IEEE/RSJ Int. Conf. Intell. Robots and Syst.* IEEE, 2012, pp. 2147–2152.

[40] 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 Int. Conf. Robot. and Biomimetics (ROBIO)*. IEEE, 2016, pp. 1948–1953.

[41] B. S. B. Dewantara, R. W. A. Saputra, and D. Pramadihanto, "Estimating human body orientation from image depth data and its implementation," *Mach. Vis. and Appl.*, vol. 33, no. 3, 2022, Art. No. 38.

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

[43] T. Ji, L. Liu, W. Zhu, J. Wei, and S. Wei, "Fast and efficient integration of human upper-body detection and orientation estimation in RGB-D video," in *2017 IEEE 9th Int. Conf. Commun. Softw. and Netw. (ICCSN)*. IEEE, 2017, pp. 1178–1181.

[44] Y. Fu and T. S. Huang, "Graph embedded analysis for head pose estimation," in *7th Int. Conf. Autom. Face and Gesture Recognit. (FGR06)*. IEEE, 2006, pp. 3–8, doi: 10.1109/FGR.2006.60.

[45] V. N. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *2007 IEEE Conf. Comput. Vis. and Pattern Recognit.* IEEE, 2007, pp. 1–7, doi: 10.1109/CVPR.2007.383280.

[46] L. Zhao, G. Pingali, and I. Carlbom, "Real-time head orientation estimation using neural networks," in *Proc. Int. Conf. Image Process.*, vol. 1. IEEE, 2002, doi: 10.1109/ICIP.2002.1038018.

[47] 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.

[48] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *Asian Conf. Comput. Vis.* Springer, 2014, pp. 82–96.

[49] B. Ahn, D.-G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robot. and Auton. Syst.*, vol. 103, pp. 1–12, 2018, doi: 10.1016/j.robot.2018.01.005.

[50] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2018.

[51] Z. Hu, Y. Xing, C. Lv, P. Hang, and J. Liu, "Deep convolutional neural network-based Bernoulli heatmap for head pose estimation," *Neurocomputing*, vol. 436, pp. 198–209, 2021.

[52] H. Liu, H. Nie, Z. Zhang, and Y.-F. Li, "Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction," *Neurocomputing*, vol. 433, pp. 310–322, 2021.

[53] M. Xin, S. Mo, and Y. Lin, "EVA-GCN: Head pose estimation based on graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2021, pp. 1462–1471.

[54] A. F. Abate, P. Barra, C. Pero, and M. Tucci, "Head pose estimation by regression algorithm," *Pattern Recognit. Lett.*, vol. 140, pp. 179–185, 2020.

[55] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: An approach for best exemplar selection," *IEEE Tran. Image Process.*, vol. 29, pp. 5457–5468, 2020.

[56] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2014, pp. 1867–1874.

[57] S.-N. Yao and C.-W. Huang, "Head-pose estimation based on lateral canthus localizations in 2-d images," *IEEE Tran. Human-Mach. Syst.*, pp. 202–213, 2024, doi: 10.1109/THMS.2024.3351138.

[58] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," in *Int. Eval. Workshop Classification of Events, Activities and Relationships*. Springer, 2006, pp. 299–304.

[59] A. K Rajagopal, R. Subramanian, R. L. Vieriu, E. Ricci, O. Lanz, K. Ramakrishnan, and N. Sebe, "An adaptation framework for head-pose classification in dynamic multi-view scenarios," in *Asian Conf. Comput. Vis.* Springer, 2012, pp. 652–666.

[60] C. Hong, J. Yu, J. Zhang, X. Jin, and K.-H. Lee, "Multimodal face-pose estimation with multitask manifold deep learning," *IEEE Trans. Ind. Inform.*, vol. 15, no. 7, pp. 3952–3961, 2018.

[61] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe, "No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, December 2013, pp. 1177–1184, doi: 10.1109/ICCV.2013.150.

[62] R. Berral-Soler, F. J. Madrid-Cuevas, R. Muñoz-Salinas, and M. J. Marín-Jiménez, "RealHePoNet: a robust single-stage convnet for head pose estimation in the wild," *Neural Comput. and Appl.*, vol. 33, no. 13, pp. 7673–7689, 2021.

[63] S. Malassiotis and M. G. Strintzis, "Robust real-time 3D head pose estimation from range data," *Pattern Recognit.*, vol. 38, no. 8, pp. 1153–1165, 2005.

[64] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3D head pose estimation using the Kinect," in *2011 Int. Conf. Wireless Commun. and Signal Process. (WCSP)*. IEEE, 2011, pp. 1–4, doi: 10.1109/WCSP.2011.6096866.

[65] R. S. Ghiass, O. Arandjelović, and D. Laurendeau, "Highly accurate and fully automatic head pose estimation from a low quality consumer-level RGB-D sensor," in *Proc. 2nd Workshop Comput. Models of Social Interact.: Human-Comput.-Media Commun.*, 2015, pp. 25–34.

[66] M. Martin, F. Van De Camp, and R. Stiefelhagen, "Real time head model creation and head pose estimation on consumer depth cameras," in *2014 2nd Int. Conf. 3D Vis.*, vol. 1. IEEE, 2014, pp. 641–648.

[67] Y. Xu, C. Jung, and Y. Chang, "Head pose estimation using deep neural networks and 3D point clouds," *Pattern Recognit.*, vol. 121, p. 108210, 2022.

[68] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *2020 IEEE Intell. Vehicles Symp. (IV)*. IEEE, 2020, pp. 1176–1182.

[69] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Inf. Process. Syst.*, vol. 30, pp. 5105–5114, 2017.

[70] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Trans. Intell. Transp. Syst.*, pp. 8063–8076, 2021.

[71] M. Chita-Tegmark, "Social attention in asd: A review and meta-analysis of eye-tracking studies," *Research in Develop. Disabilities*, vol. 48, pp. 79–93, 2016.

[72] S. Clifford and C. Dissanayake, "Dyadic and triadic behaviours in infancy as precursors to later social responsiveness in young children with autistic disorder," *Journal of Autism and Develop. Disorders*, vol. 39, no. 10, pp. 1369–1380, 2009.

[73] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, "Early social attention impairments in autism: social orienting, joint attention, and attention to distress." *Develop. Psychology*, vol. 40, no. 2, p. 271, 2004.

[74] W. Jarrold, P. Mundy, M. Gwaltney, J. Bailenson, N. Hatt, N. McIntyre, K. Kim, M. Solomon, S. Novotny, and L. Swain, "Social attention in a virtual public speaking task in higher functioning children with autism," *Autism Research*, vol. 6, no. 5, pp. 393–410, 2013.

[75] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, J. Lu, Q. Peng, and X. Qu, "Atypical head movement during face-to-face interaction in children with autism spectrum disorder," *Autism Research*, vol. 14, no. 6, pp. 1197–1208, 2021.

[76] Z. Zhao, Z. Zhu, X. Zhang, H. Tang, J. Xing, X. Hu, J. Lu, and X. Qu, "Identifying autism with head movement features by implementing machine learning algorithms," *Journal of Autism and Develop. Disorders*, pp. 1–12, 2021, doi: 10.1007/s10803-021-05179-2.

[77] M. Argyle and M. Cook, *Gaze and mutual gaze.* Cambridge University Press, Cambridge, UK., 1976.

[78] R. Vertegaal, R. Slagter, G. Van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes," in *Proc. SIGCHI Conf. Human Factors in Comput. Syst.*, 2001, pp. 301–308.

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

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

[81] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Tran. Pattern Anal. and Mach. Intell.*, vol. 21, no. 5, pp. 476–480, 1999.