Measuring inferred gaze direction to support analysis of people in a meeting

Authors
Robert Wright
Email: K1448941@kingston.co.uk

Tim Ellis
Email: T.Ellis@kingston.ac.uk

Dimitrios Makris
Email: D.Makris@kingston.ac.uk

Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road Campus, Kingston upon Thames, KT1 2EE

Corresponding Author Details
Robert Wright
Email: K1448941@kingston.co.uk
Mobile: +44 (0) 7885613462
Postal Address: 85 Crossways, Gidea Park, Romford, Essex, RM2 6AS, England

Declarations of interest: none

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

This paper introduces a method to infer gaze direction and the point of focus of participants involved in a collaborative activity, such as a meeting. It uses a single depth-based sensor placed overhead to capture the meeting, which has the benefit of avoiding occlusion and is unobtrusive, minimising possible changes in behaviour that might arise if people are aware of the sensor. The inferred gaze direction of each participant is estimated in the horizontal plane from the orientation of the head (yaw), derived from a segmentation of the depth image to generate an outline of the head. A common focus of attention is inferred by intersecting the gaze directions of each participant. Performance evaluation using a depth camera to record a meeting achieved a head detection performance of 99.6% and a valid gaze detection of 96.9%.

Keywords

Gaze estimation; Visual Focus of Attention; gaze intersection; Depth capture; group meeting analysis

1. Introduction

Analysing the interaction between people has gained increasing interest in recent years with the desire to understand the dynamics of people in collaborative settings, such as meetings and team discussions. Typical interactions include following a conversation or discussion, watching a presentation, or interacting with an object in the scene, focusing on one another or the inevitable

distractions from the surrounding environment. The complexity of this task increases with the number of participants.

A variety of technologies have been used to capture different aspects of this interaction, including microphones, cameras and depth sensors. However, if people are aware of the technology that is observing them, or are asked to wear the technology, it can change their behaviour, a reaction that is referred to as the observer or Hawthorne effect (Roethlisberger et al., 1939). Alternatively, an unobtrusive approach restricts many of the current ways in which data is gathered in meeting scenarios. Analysis of the interaction of a group of people who are collaborating provides a significant challenge, not only to capture all of the people together but also to detect the interaction that takes place between them and with other elements in the scene.

Previous work on this challenge (e.g. Kaminski et al., 2006) tries to capture the interaction by detecting each person's visual focus of attention (VFOA) using a variety of alternative sensors (e.g. cameras, eye trackers, orientation sensors, sound source location). A widely used approach (Kaminski et al., 2006) is to capture the head pose and facial features of individuals to determine the orientation of the head and infer the direction of gaze. Gaze direction is composed of two elements – the orientation of the head and eye orientation. In searching for a new fixation point, it is normal for the eyes to move first, followed by a head rotation (and a compensatory eye rotation in the opposite direction). The eyes have a horizontal ocular range of approximately +/- 25° (Stahl, 1999), and outside of this range a change of head orientation is needed for fixation. Cameras may be used to detect the eyes and directly extract the direction of gaze; in addition, if both eyes can be observed it may also be possible to estimate the depth of fixation.

However, except in constrained circumstances, a single camera is insufficient for a group of people and multiple cameras need to be employed. Other approaches (Masse, Horaud, 2017) have used elaborate head-mounted equipment with a sensor to measure both head orientation and also eye direction. Depth cameras have also been used to determine the direction of gaze by establishing the shape or pattern of the head from the depth information to determine the orientation of the head (Bhattacharya et al., 2018).

Many of these technologies are intrusive and may influence a person's natural behaviour, adversely impacting the measurement. Increasing the number of capture devices complicates the installation and usage of the system and increases the complexity of the image analysis, with potential problems of occlusion and the integration of measurements from each camera. Also, it is necessary to globally calibrate such person-centred measurement systems so that the interaction between individuals can be determined.

As well as estimating the gaze direction of each individual, additional information can be gained by identifying the periods when the attention of several (or all) of the group is focussed on a specific target, such as the person who is speaking, or an object of interest that is the subject of the discussion. In this case the collective attention of the group may be inferred by computing the intersection of the individual gaze directions. Whilst such collective gaze behaviour is common it is by no means consistent, and the gaze of individuals can be subject to distraction or observing non-speakers to observe reactions.

This research aims to provide a methodology to support the analysis of interaction within a collaborative setting, using an unobtrusive approach with minimal impact on the natural behaviour of the participants, thus enabling an effective interpretation of the actions and interaction of the participants. Our approach is to infer the gaze direction of each participant by estimating the

orientation of the head in the horizontal plane (yaw), derived from a segmentation of the depth image to generate an outline of the head using a single overhead sensor. A particular challenge for this approach is that, since the eyes are not visible, the true gaze cannot be directly measured. A specific aim of the research is to investigate the capabilities and limitations of the method under this (and other) constraints.

The novelty of the research lies in the following: the use of a single depth sensor, placed overhead; detection of participants from maximally-stable extremal regions; modelling the gaze with a Gaussian distribution; and computing a common fixation point that accommodates inaccuracy in the gaze direction estimates. Throughout the text the term "gaze direction" is used to refer to this inferred gaze direction; where the true gaze (i.e. where the eyes are pointing) is referenced the term "true gaze" will be used.

Our unobtrusive approach is applied to recordings of a group of people sitting around a meeting table. In calibrating the scenario, the participants are asked to carry out a controlled activity by fixating on a specific object viewed at different locations; a second interactive activity used an object passed around the group as the subject of a discussion. Our approach will leverage the technique of overhead data capture, maximising the benefits of an unobstructive view of the participants and their interactions. The inferred gaze directions of the participants are used to Identify the subject of groups focus of attention, whether that is an object, an activity, or the speaker.

This paper focuses on extracting measurements of people in a meeting, in order to support behavioural studies on group focus and attention, such as the gaze change in the activity of turn-taking in discussion (Ho et al., 2015). Other applications may be in teaching, where children's FOA is measured for levels of engagement or focused on one-on-one teaching in cases of special needs; or a person performing a specific activity, such as playing an instrument.

2. Related work

In humans, the direction of someone's gaze provides insight into their focus of interest (Mareschal et al., 2013). Gaze also plays a part as a cue within collaborative settings, confirming focus or attention to the speaker (Stahl, 1999, Kluttz et al., 2009, Wilson et al., 2000, Daar et al., 2012). The review of the different gaze direction capture methodologies uncovered a need to understand the relationship between the eye and the head, how they work together and the natural process of gaze and fixation within the activities measured.

(Stahl, 1999) considered the interaction between head and eye motion in the horizontal plane, associating large changes in gaze direction with head motion, while saccades support smaller changes. He noted that the head catches up with the eyes, and the final configuration is with the eyes looking straight to the front.

Research to capture gaze direction and VFOA fall mainly into two approaches. Firstly, a combination of head pose and eye movement with the technology attached too or directly in front of the person (Kaminski et al., 2006, Ba et al., 2006, Ghiass et al., 2016, Fischer et al., 2018). Secondly, capturing head orientation and inferred gaze estimation from multiple cameras overhead or a central omnidirectional camera on a table, providing an opportunity to observe more than one person simultaneously (Tian et al., 2003, Cohen, et al., 2000, Wu et al., 2017).

Research conducted by (Chong et al., 2018) presented a method of identifying gaze and attention of subjects looking at objects, at the camera or out-of-frame gaze targets from images. Chong et al. developed a generalised visual attention estimation method which learns a subject's saliency map and a three-dimensional vector from the yaw and pitch of the heads in the images. Similar

techniques of facial models (Kaminski et al., 2006), used anthropometric features, in this case, live tracking of the four corners of the eyes and the point of the nose.

An alternative approach (Voit, Stiefelhagen, 2008, Voit, Stiefelhagen, 2010) used four cameras to track in 3D the head orientation of participants in a dynamic meeting scenario. This approach coped with occlusions and removes the need for capturing facial features to estimate gaze direction. By predetermining possible points of interest within the space, the system was able to predict the individual participant's focus of attention in up to 72.2% of all frames.

In a variation to this approach to capture more than one person at a time, (Stiefelhagen, 2002) used an omnidirectional camera and microphones, the later to support direction analysis, positioned in the centre of the meeting table, capturing the head orientation of all participants. Microphones to support location measurements of gaze fixation points were also used in (Reddy, 2016) as part of an overhead RGB camera configuration to capture head orientation. The approach of (Stiefelhagen, 2002, Reddy, 2016) combined visual and sound analysis and managed to achieve 75.6% and 70% accuracy respectively for identifying the focus of attention, i.e., the person talking in both experiments.

There is the assumption that when people interact, they tend to look in the direction of the relevant object or person. (Masse, Horaud, 2017) used the head orientation and the eye gaze direction to identify the intended focus of attention on people and objects, recognising that people look at a person who is talking or at an object of interest under discussion. In tracking gaze direction in three or more participants, the capture of eye gaze is difficult, as eyes become barely detectable as participants turn to face each other, and requires multiple cameras. In the method deployed by (Masse, Horaud, 2017) a Bayesian switching dynamic model used the orientation or head pose to track the gaze direction. When both eyes were in view, a trackable learning algorithm further increased the accuracy of determining the focus of attention.

(Hu, G. et al., 2014) focused on tracking human activity using a top-down three-dimensional camera approach. Their hierarchical tracking models scene constraints and motion patterns to find and track people in a space used for action recognition, applying the method to tabletop activity research and vertical screen interaction measurement. This approach removes occlusion despite losing facial information, but with many human movements and actions still distinguishable.

(Reddy, 2016) adopted a similar top-down approach to (Cohen et al., 2000), using RGB images from an overhead camera to capture the vocal interaction of a group of people. By intersecting the gaze direction of each participant, the speaker could be identified 70% of the time. (Bhattacharya et al., 2018) described a multimodal conference room that uses two types of depth sensors positioned overhead and microphones attached to each participant in their study of human dynamics. The main focus of this research is centred on non-verbal speech patterns and discussion content to identify group rankings of possible leaders and major contributors. The use of overhead sensors to estimate participant's head pose and VFOA as with (Reddy, 2016, Wu et al., 2017, Bhattacharya et al., 2018, Tian et al., 2003, Cohen et al., 2000), demonstrates a growing trend for overhead measurement.

Each of the approaches described above present challenges: for example, it is difficult to use a single sensor to capture facial features (and to infer the direction of gaze) of more than one person. Also, with only a single sensor, problems occur due to occlusion unless the participants are arranged in a very constrained configuration. The use of an overhead sensor (Reddy, 2016, Wu et al., 2017, Cohen et al., 2000, Bhattacharya et al., 2018) provides a clear sight of the participants and minimises the

possibility of occlusion. The experiment set out by (Hadjakos, 2012) to track head and hands suggest the possibilities for other applications of capturing human activities for study. Although not directly focused on gaze direction, it does provide insight into the methodology taken by this research.

Placing the capture device directly overhead addresses several issues associated with viewing faces from a frontal position. Firstly, by minimising the problem of occlusion, providing a clear view of the entire scene of interaction; secondly, it avoids the problem of the capture device located in the line of sight of the participants, reducing the possibility of affecting the natural behaviour of the participants. Thirdly, using a single capture device avoids the complication of needing to combine the measurements from multiple sensors, which would require geometric calibration to determine their spatial arrangement. The single sensor approach is simpler to set up and more easily repeatable.

The choice of a depth sensor over the more conventional RGB camera has clear benefits: the heads of the individuals are distinctly and unambiguously visible in the depth map and are easily separated from other parts of the body and other elements in the scene, simplifying the segmentation task, whereas the segmentation of an RGB image must cope with the variability of varying quantities of hair, as well as hair colour; lastly, using depth data supports the removal of personally identifiable information from data sets, making it viable for use in office spaces.

The novelty of this paper is the use of a single depth-based sensor to capture the inferred gaze direction (in the horizontal plane), of multiple interacting participants. It differs from previous work by providing a clear field of view of the participants and their interaction with objects. It significantly simplifies head detection by making novel use of the MSER (Maximally Stable Extremal Regions) algorithm to slice the depth data through each head, which yields an elliptical shape whose orientation is used to estimate the gaze direction. Intersecting individual gazes modelled by Gaussian distributions enables detection of the collective attention focus, whereas many of the works described above simply treat the gaze vector as a line and ignore the uncertainty of the measurement.

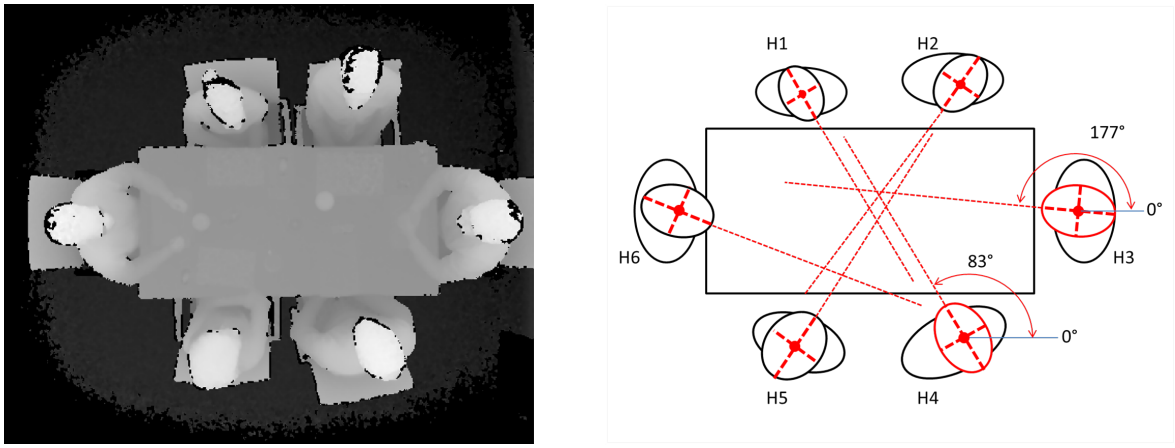
3. Methodology

The scenario where we undertake the analysis for this research is for a group of participants engaged in a conventional meeting, spaced around a table (see figure 1). The principal assumption is that the people being observed are in an upright posture so that the head is clearly visible. As shown in the diagrammatic representation in Fig. 1b, the gaze directions of participants can be used to infer their VFOA, and particular events (e.g. someone speaking, an object of interest) can be detected by determining their collective VFOA's.

The approach uses a single unobtrusive sensor mounted overhead, simplifying the measurement process by limiting the gaze to the horizontal plane (yaw) as it would be more difficult to extract reliable estimates of head pitch with this imaging geometry. However, the measurement of yaw is the most useful axis for the subsequent analysis in detecting the common fixation point for the scenario we are using. The overhead sensor provides compact and self-contained imaging of the scene, which simplifies the computation of gaze detection and common focus point, but imposes a constraint because the eyes are not visible and determining the pitch of the head from this viewpoint would be unreliable and subject to inaccuracy.

There are three key steps in this methodology: firstly, detecting the heads of the N participants; secondly, determining the gaze direction of each participant; and thirdly, identifying potential fixation points for each individual. A transverse section through the human head can be

approximated by an ellipse. Depth data captured from an overhead sensor are pre-processed to minimise noise. Next, heads are detected as maximally stable extremal regions and represented by the best-fitting ellipse. The N regions closest to the sensor are considered valid heads of the N participants, and are associated on a frame-by-frame basis using proximity. The gaze direction of each head is derived from the orientation of the major axis. The gaze direction is intersected with each of the other participant's head locations (centroids) and any common fixation point is found from the accumulation of intersections.



(a) (b)

Figure 1. a) Overhead depth image, b) head detection, inferred gaze direction and common focus point where multiple gaze lines intersect.

3.1 Head Detection and Representation

Depth images are noisy and require pre-processing before detection. Fig. 1a shows a frame from an overhead depth image of six people sitting around a meeting table. The black pixel areas around the edges of the heads and bodies correspond to locations where the sensor has failed to measure the depth for the pixel and returned a zero value. Such zero-depth measurements happen because some surfaces and materials (such as hair) are not strong reflectors of the infrared illumination used to acquire the depth image. This can be seen in Fig. 1a around the boundaries of each head, resulting in a distortion of the size and shape after segmentation.

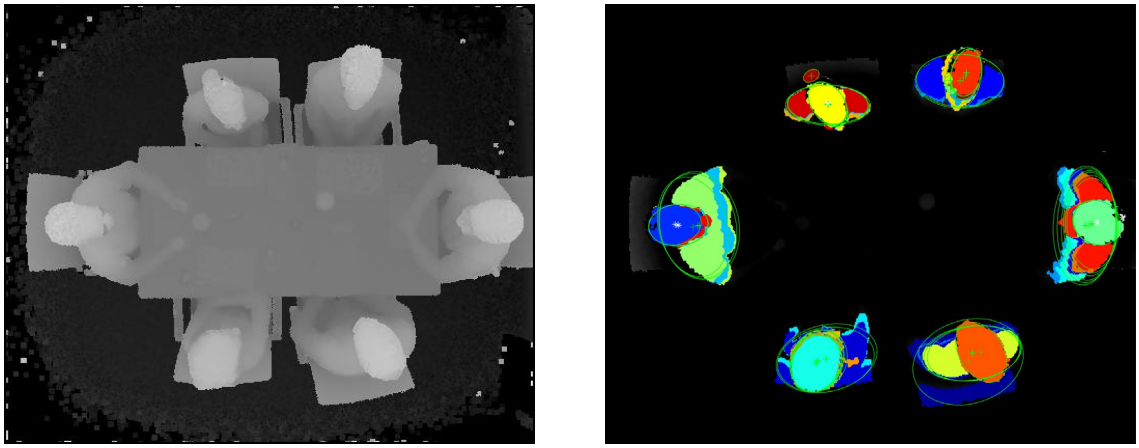
Pre-processing is carried out on each image, based on clipping the depth values into an 8-bit range to eliminate any readings that are greater than the actual depth between the sensor and the floor; additionally, the depth information is inverted to provide depth values that range from a low value at the ground to a high value at the top of the heads (and hence the heads appear brighter when displayed). Finally, a 5x5 non-linear Minimum filter is used to suppress the black pixels around the heads. The filter is modified so that it is only applied at locations where the centre pixel value is zero; in addition, only the non-zero values in the 5x5 window are evaluated to determine the minimum value which is used to replace the zero centre value. The result of applying this filter to the image shown in Fig. 1a can be seen in Fig. 2a, where the black pixels around the heads have been clearly suppressed.

The head of each participant is detected by thresholding the pre-processed depth image using the MSER (Maximally Stable Extremal Regions) algorithm (Matas et al., 2004), which segments the depth image into non-overlapping regions. The MSER algorithm identifies connected regions (R) of pixels with an adjacency (connectivity) relationship, using an independent threshold for each region that is

computed as the algorithm descends through the depth values in each region. The algorithm uses two parameters to segment the image into regions: the incremental step size, which determines by how much the threshold changes at each iteration; and the maximum change in area between one binary slice and the next, which determines the stability of the segmentation.

A problem may arise if the participants exhibit a very wide range of heights, where the shoulder height of one individual exceeds the head height of the shortest person, such as a child and a basketball player. In this situation the N MSER regions closest to the camera will miss the head of the smaller person. Such a situation is addressed by considering the centroids of each of the selected regions as the centroids of the head and shoulders are largely coincident, and a simple test on the distance between the selected region centroids allows the shoulder region to be rejected.

Fig. 2a shows the result of the depth image after pre-processing is applied. Fig. 2b shows the multiple binary regions that are returned by the MSER algorithm and constructions of the best-fitting ellipses.



(a) (b)

Figure 2. (a) Depth image after pre-processing, (b) segmented regions, resulting from applying the MSER algorithm; including best-fitting ellipses and centroids.

Each region is then represented by its zero, first and second order moments (corresponding to the area (in pixels), centroid location and the major and minor axis and orientation of the best fitting ellipse) and its mean depth. The mean depth computed for each region is used to rank the regions in the order of their distance from the camera. The first N regions of this sorted list are taken to represent the detection of N participant heads.

The constrained and largely static setup of the meeting table means that the task of tracking each individual is a trivial step of data association from one frame to the next by computing the Euclidean distance between the centroids of every pair of heads, and uniquely associating those with the minimum distance.

The output of the detection stage is N regions. Each region $(R_n, n = 1..N)$ is described by an elliptical model represented by a centroid (x, y), major and minor axis a, b and the angle θ, the major axis of the fitted ellipse

$$R_n = \{x_n, y_n, a_n, b_n, \theta_n\} \quad (1)$$

3.2 Gaze Direction

The gaze direction of each person is inferred from the head orientation as the eyes are not visible, based on the orientation θ. To address the 180° ambiguity in the orientation in determining the gaze direction, a constraint is imposed that the participants are looking in the direction towards the centre of the table (rather the away from the table centre). The uncertainty of the gaze direction (GD) for each participant is modelled by a Gaussian distribution, $GD(\mu,\sigma)$, constructed using estimates of the mean gaze direction and its standard deviation (σ_d) based on ground truth measurements:

$$GD(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

This 1D distribution is projected along the direction of gaze from the head centroid $c(x, y)$ using the standard deviation of the gaze direction to create a 2D 'fan' distribution. The value of σ varies linearly along the gaze projection ray and is computed from $\sigma = s \tan \sigma_d$, where s is the distance along the ray from the head centroid. Each 1D distribution is normalised so that the peak has a value of 1, to ensure that the weight along the most likely direction of gaze is independent of the distance from the viewer. Fig. 3a shows a set of the Gaussian profiles forming the 2D fan distribution; Fig. 3b shows the 2D distribution of the gaze projection model in image space, originating at a head centroid.

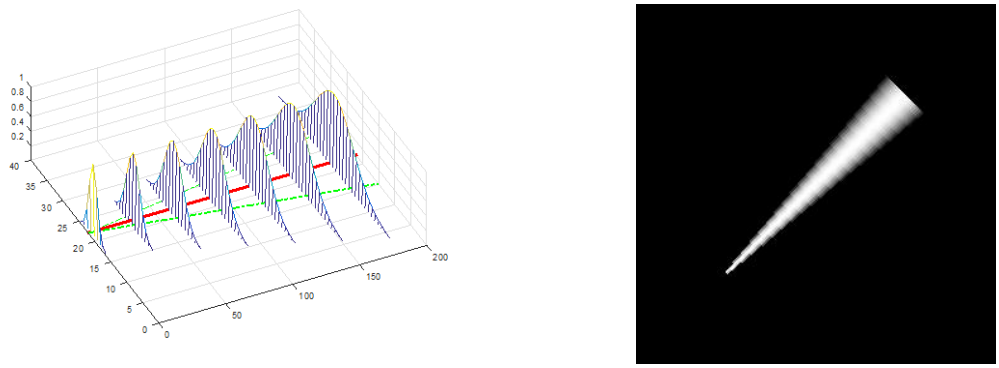


Figure 3. a) Projection of Gaussian functions along gaze direction (red line) and the standard deviation (green lines); b) normalised 2D Gaussian fan distribution projected into image space. The image has been logarithmically enhanced.

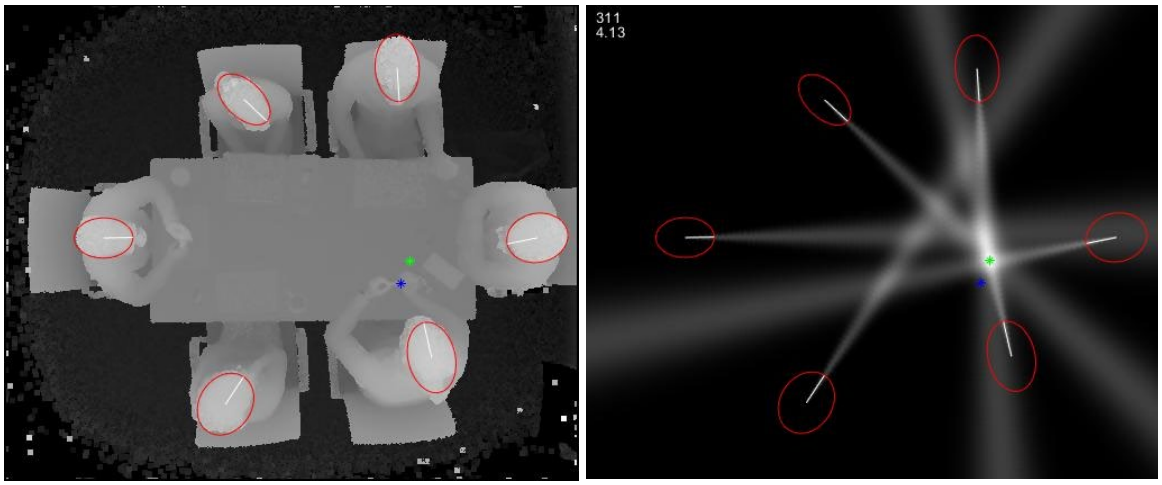
The gaze direction of each person is analysed to determine what they are observing (or looking towards) in the scene. Likely "targets" of their observation are the other participants in the meeting (e.g. the person speaking), or objects of interest that are the subject of discussion. By detecting the coincidence or intersection of the participant's gaze directions, we can infer the detection of a discrete object or a particular speaker.

To detect the most likely target an individual may be observing, the minimum distance between the observers projected gaze and the centroid of the other five heads is computed. The observed target is assumed to be the one with the smallest distance. If the object is being observed, its location can only be estimated by triangulation of two or more gazes that is described as a common fixation point.

3.3 Common Fixation Point (CFP) Detection

The CFP is computed using a maximum likelihood-type approach and implemented by a Hough-like spatial accumulator of the Gaussian gaze direction models. The fan distribution of the gaze direction of each participant projects into the accumulator space forming intersections, with detection identified as peaks in this space. By normalising each model so that the peak value is one along the gaze direction, the weight of votes in accumulator space corresponds to the number of participants looking towards the same location in space and are determined from the location of the maximum value, and hence the most likely location of the CFP.

For a group of N people, the peak magnitude for all participants observing the same location would be a value of N, though this would require all their gazes to be directed exactly at the target (note: this would actually be N-1 if the FOA was a speaker). The location of the largest cluster is used to infer a CFP shared by the participants. Fig. 4b shows the accumulator space of the image shown in Fig. 4a. The green star marks the location of the largest peak detected in the accumulator space; the blue star corresponds to the known location of the object that is being discussed (and held by participant H4, as labelled in Fig. 1b).



(a) (b)

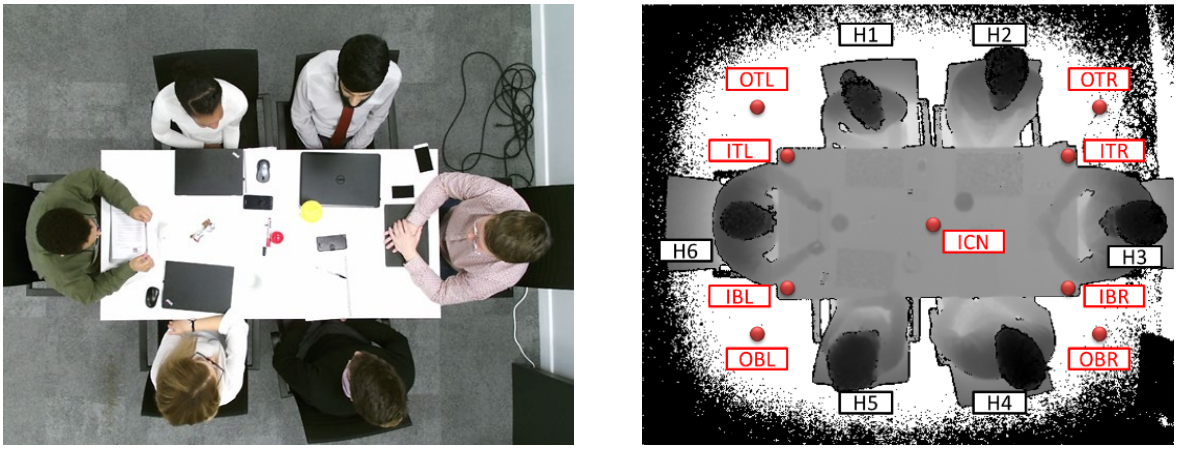
Figure 4. a) Detection of heads with their elliptical approximation (red ellipses) and estimated gaze direction (white lines). The known location of the object of interest, marked by a blue star and the CFP by a green star. b) The plot of the accumulator space of gaze directions; the magnitude of the largest peak is 4.13.

4. Data Collection

4.1 Experiment Setup

A Kinect V2 sensor was used to capture depth data from the scene. The sensor was placed 2.8m above the floor which is consistent to the recommendation given by (Wu et al., 2017), avoids any potential data degration expected beyond 4.5m distance and allows sufficiently large FOV and accuracy of measurement between the ground and the head locations of the participants.

Each frame of depth data provides an output of 512 x 424 pixels. Fig. 5a shows the output from the Kinect RGB camera; Fig. 5b shows the associated depth image (approximately aligned). The capture rate was an average of approximately 2fps for the two experiments described below. The Matlab implementation of the MSER algorithm was applied using a step size of 1.75 and a maximum area variation between extremal regions of 0.4 to produce the most reliable results (see Section 5.1).



(a) (b)

Figure 5. a) RGB image of scene, b) depth image showing labelled locations of the red ball used in Experiment 1.

Data were captured to enable completion of two activities, a "controlled object focus activity" where participants were asked to fixate on a fixed point at different locations around the table. The second activity recorded "a discussion that involved interacting with an object" to evaluate the methodology applied to a real task.

4.2 Experiment 1 – Comparison with Ground Truth

The first experiment assesses the accuracy of the algorithm to estimate the gaze direction using manually acquired ground truth, based on the depth image (see figure 2). Manual measurement involved determining the coordinates of the centre (centroid) of the head and estimating a point in front of the head in the estimated direction of gaze. The orientation of the line joining these two points is taken as ground truth. The repeatability of the manual measurement was evaluated by making multiple measurements on a set of heads. The head orientation obtained from the ground truth measurements is compared with the gaze direction computed from the ellipse orientation.

4.3 Experiment 2 - Controlled object focus activity

The purpose of this experiment was to assess the reliability of estimating a person's gaze direction from a measurement of the orientation of the head, knowing that the true gaze is determined by a combination of head and eye orientation. The aim to ask the participants to look at a distinctive target placed at locations in the scene, and then compare the computed gaze with the known location of the target. In face to face interaction this may not be a natural use of gaze as people tend to look away and back when the person you are conversing with is speaking (Ho et al., 2015). For this exercise we relied on the participants following the instructions for this experiment and to maintain their gaze on the object.

This activity used a red ball as a point of focus for the participants. Fig. 5b shows the red ball in nine separate locations, five locations on the surface of the table and four locations off the table surface where the ball was held in position by a seventh person. The participants were asked to focus on the red ball for 5 seconds at each location, although in some cases the number of recorded frames is less. The results of the computed gaze direction were compared with the angle from the head centroid to the known location of the red ball.

4.4 Experiment 3- Interacting with an object activity

The purpose of this experiment is to quantify the level of attention of each of the participants to a specific activity. For this task, the participants were asked to interact and discuss an object; the

object was unusual and was detailed in its manufacture to encourage people to focus closely on it. Participants were not asked to hold the object in any specific way, but instead, they were left to view the object naturally and only told to pass it between each other so all could take part in the activity. Figure 6 shows 9 locations where the object was approximately stationary (as each participant manipulated the object) are represented by the circular regions depicted in Fig. 6, which were manually measured.

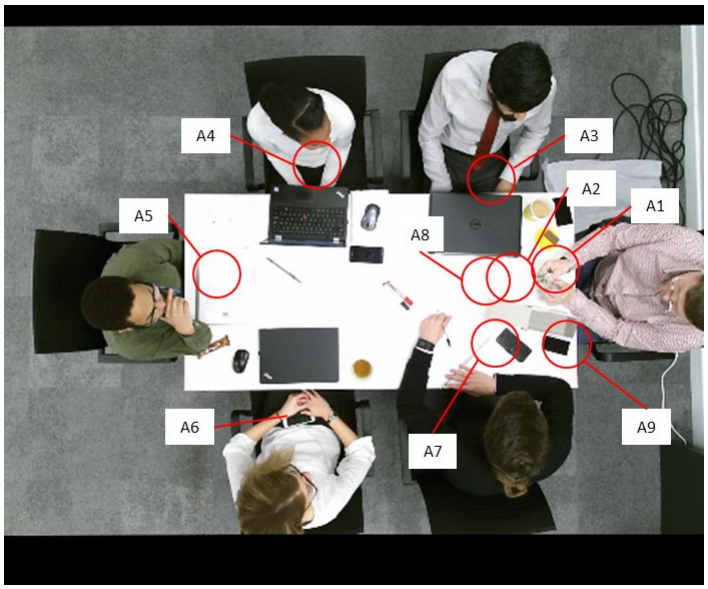


Figure 6. Red circles identify the nine approximate locations (A1-A9) of the object location throughout the session

5. Results and Discussion

5.1 Head Detection

The implementation of the MSER algorithm uses two parameters to segment the image into regions: the step size and the maximum area variation. We evaluated the performance of our head detection method on 3000 depth frames sampled from six data sets (fig. 7 below) for a range of the parameter values. For the step size, a range of 1.50 to 2.50 (around the default value of 2.0 used in the Matlab implementation) at intervals of 0.1 was tested, with the maximum area variation fixed at 0.4. The best performance for detection was found to be between 1.60 and 1.90, with a drop off in detection performance beyond 2.3. The maximum area variation parameter was evaluated, using a step size set to 1.75, over a range of 0.2 to 0.8 at intervals of 0.05 (fig. 8b below). The best performance was found over the range 0.30-0.70 with a drop off in detection before 0.30. The detection rate is consistently above 98% for a wide range of values in both experiments for all datasets (Fig. 8), indicating that our head detection algorithm is not sensitive to the selection of parameters. Also, it is superior to (Hu, G. et al., 2014) that reported 95.5% detection rate. For the rest of the experiments, the step size between successive threshold levels and the maximum area variation are set to 1.75 and 0.4 respectively, based on the above results.

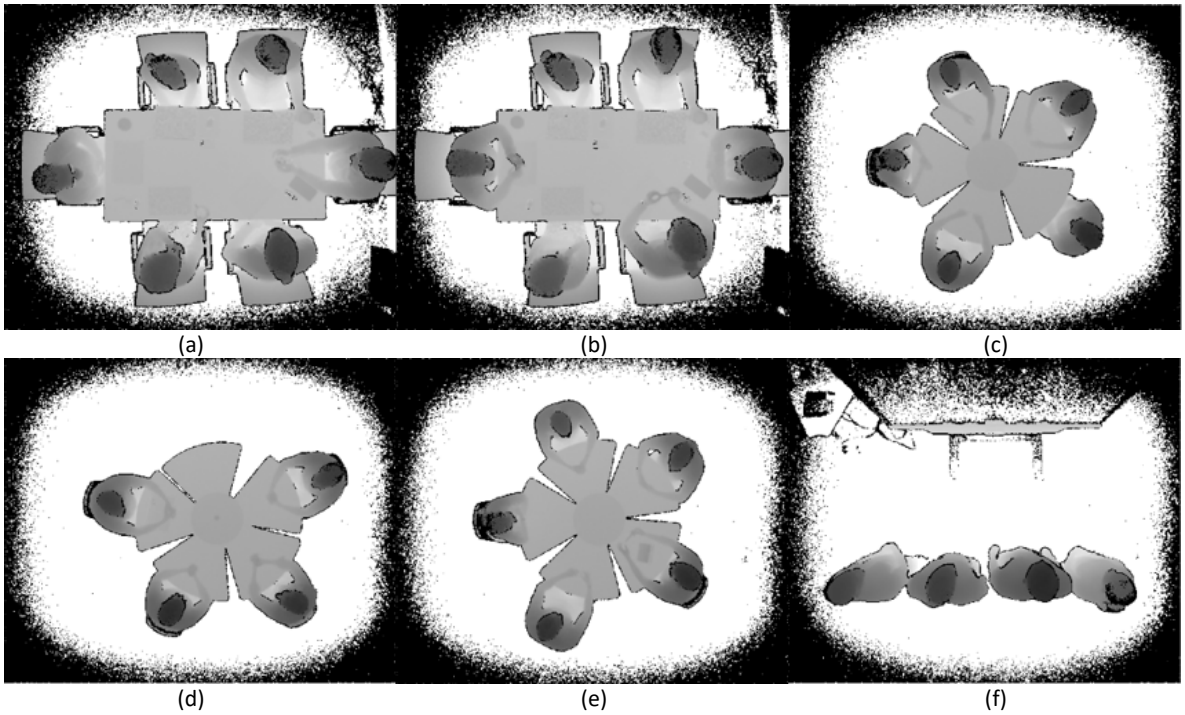


Figure 7. Example images from the six data sets of raw Kinect V2 depth images where depth values are measured as from the camera: a) 6 person object discussion (Dataset 1), b) 6 person general discussion (Dataset 2), c) 5 person discussion (Dataset 3), d) 4 person control session (Dataset 4), e) 5 person object session (Dataset 5) and f) 4 person standing group (Dataset 6).

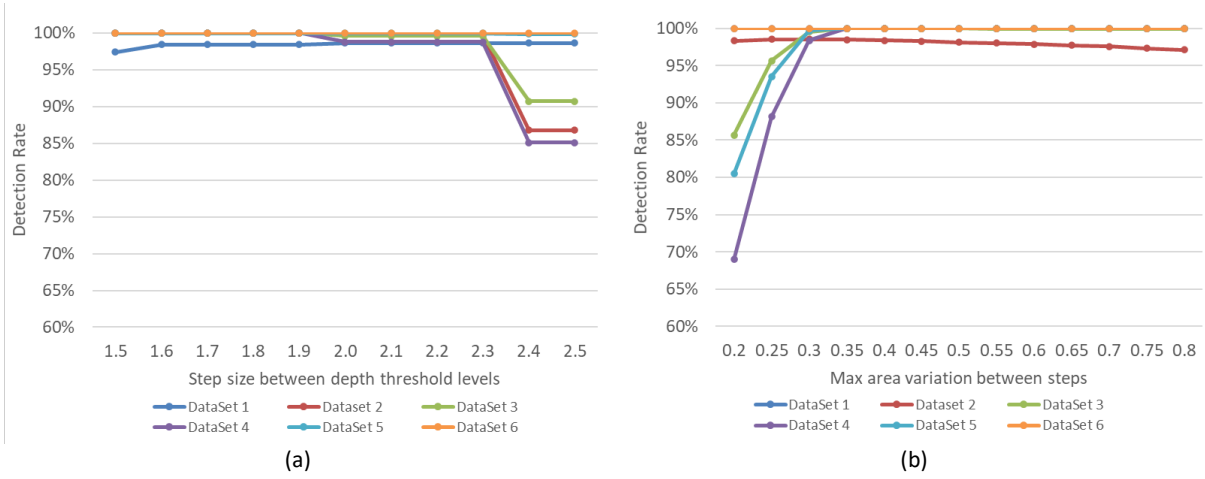


Figure 8. Head detection performance results for the MSER algorithm for varying parameter settings: a) step size and b) maximum area variation between steps.

The MSER implementation also filters out regions outside of an acceptable size by specifying an area range. This was used primarily to reject small noisy regions and was set to [200-10000] pixels for the experiments reported here. Given that the typical size of detected head regions is in the range of 1000-2000 pixels, it can be seen this is not a critical setting for the detection.

Table 1 shows the results of head detection for each participant over the 1447 frames acquired for Experiment 3; detection is confirmed when the threshold parameters of the head ellipse have been met. The overall detection accuracy is 99.94%, with only 5 instances where a head was not detected.

	H1	H2	H3	H4	H5	H6	Head count	Rate
Total frames	1,447	1,447	1,447	1,447	1,447	1,447	8,682	
Head Detected	1,444	1,447	1,447	1,447	1,447	1,445	8,677	99.94%
Head Detection Error	3	0	0	0	0	2	5	0.06%
Error Rate	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%		
Valid Gaze	1,447	1,447	1,183	1,447	1,445	1,445	8,414	96.9%
Invalid Gaze	0	0	264	0	2	2	268	3.1%
Error Rate	0.0%	0.0%	18.2%	0.0%	0.14%	0.14%		

Table 1. Head and gaze detection results for each participant over 1,447 frames.

The bottom half of Table 1 shows the results of finding a valid gaze direction for each head, based on the orientation of the best-fitting ellipse. Errors in this measure arise as a consequence of head 3 (H3) intersecting the edge of the image FOV, which resulted in incomplete detection of the head and a poorly fitting ellipse. Gazes for heads 5 and 6 (H5 and H6) were also invalid for two frames. These results indicate that the head detection algorithm is robust and provides a valid gaze, as long as the heads are within the image field-of-view. Our valid gaze detection (96.9%) is clearly superior to (Reddy, 2016) that reported 70% correct detection field

5.2 Gaze Direction

The data collected from experiment 1 are used in this section to evaluate the gaze estimation, firstly against manually estimated ground truth, and then against the known locations of the red ball.

5.2.1 Evaluation of gaze estimation against ground truth.

Accuracy and repeatability of the manual ground truth method were assessed by making 10 repeated measurements (M1 to M10) on a single depth frame. Measurements of the six heads were taken from the centre (centroid) of the head to the estimated direction of gaze with the angle generated from the two points. Table 2 shows the angles recorded from this validation exercise over the ten sets of measurements. The average standard deviation for all measurements is 1.3 degrees.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Mean	StdDev	Min	Max
Head 1	265°	264°	264°	265°	262°	263°	264°	265°	264°	264°	263.9°	0.8°	262°	265°
Head 2	231°	234°	234°	232°	234°	231°	232°	234°	230°	236°	232.6°	1.8°	230°	236°
Head 3	178°	178°	178°	179°	178°	178°	177°	178°	179°	179°	178.2°	0.7°	177°	179°
Head 4	138°	138°	136°	136°	137°	137°	137°	137°	139°	133°	136.9°	1.7°	133°	139°
Head 5	79°	78°	75°	80°	78°	79°	76°	78°	78°	76°	77.7°	1.5°	75°	80°
Head 6	347°	347°	348°	346°	347°	347°	346°	346°	344°	345°	346.3°	1.0°	344°	348°

Table 2. Results of the ground truth measurement of head orientation (in degrees) for the six heads repeated ten times on the same frame to validate the measurement method.

Table 3 shows an evaluation of the automatic gaze estimation against the ground truth, provided by manual measurement of the head orientation. Manual ground truth measurements were made for 140 image frames for each of the six heads in each image, giving 840 ground truth measurements in total.

The range of error for Heads 1, 2, 4 and 6 (H1, H2, H4, H6) show a similar low mean difference and standard deviation. Both Head 3 (H3) and Head 5 (H5) have higher mean differences: for Head 3 the

head intersects the edge of the FOV; for Head 5 the difference can be attributed to the participant's hairstyle, which it is observed can distort the measurement if it is asymmetric. Hence, we can conclude that the automatic measurement can be reliably measured from the image data with an average error of 7.7° and standard deviation of 4.9°.

	H1	H2	H3	H4	H5	H6	Average	Frames
Mean	6.8°	5.9°	12.8°	4.0°	12.9°	3.8°	7.7°	140
StdDev	4.7°	3.7°	5.0°	3.1°	7.6°	4.1°	4.9°	140

Table 3. Mean angle (in degrees) difference and standard deviation between the algorithmically estimated angle and the ground truth angle generated from the manual measurement of head orientation for 140 frames from the start of the Experiment 1 data sequence.

Fig. 9 shows a graphical comparison of the manually measured head orientation with the computer-estimated gaze direction angle for one of the participants, head 1 (H1) for 140 frames of Experiment 1 data. The average standard deviation over all the measurements is 4.9 degrees and this value is used for σ_d in computing the spread of the fan distribution. These results indicate that the head detection algorithm is fairly accurate, aligning with a manually measured ground truth results.

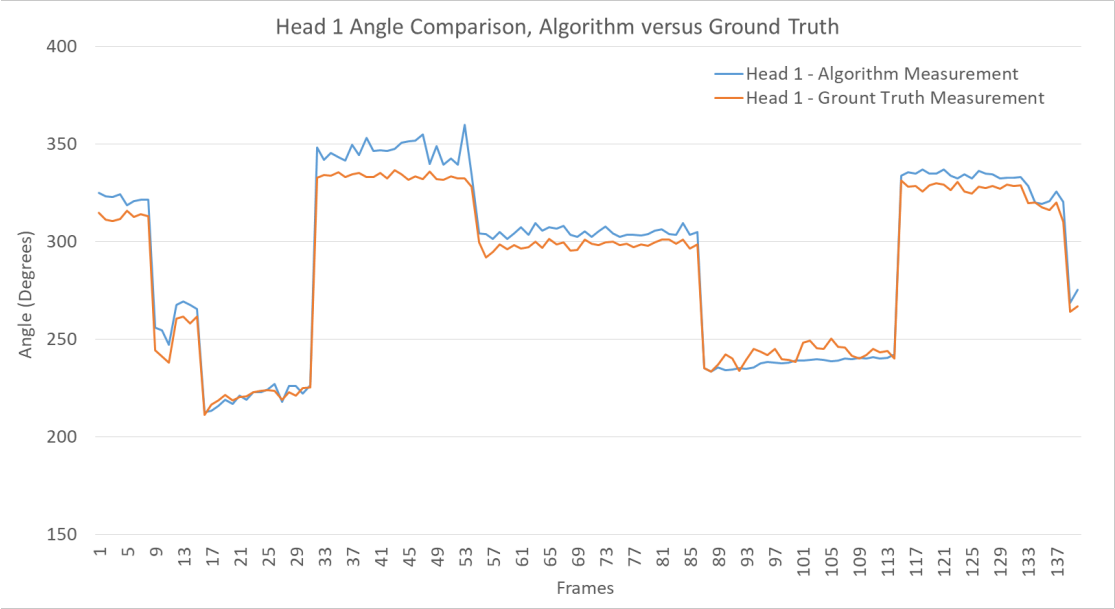
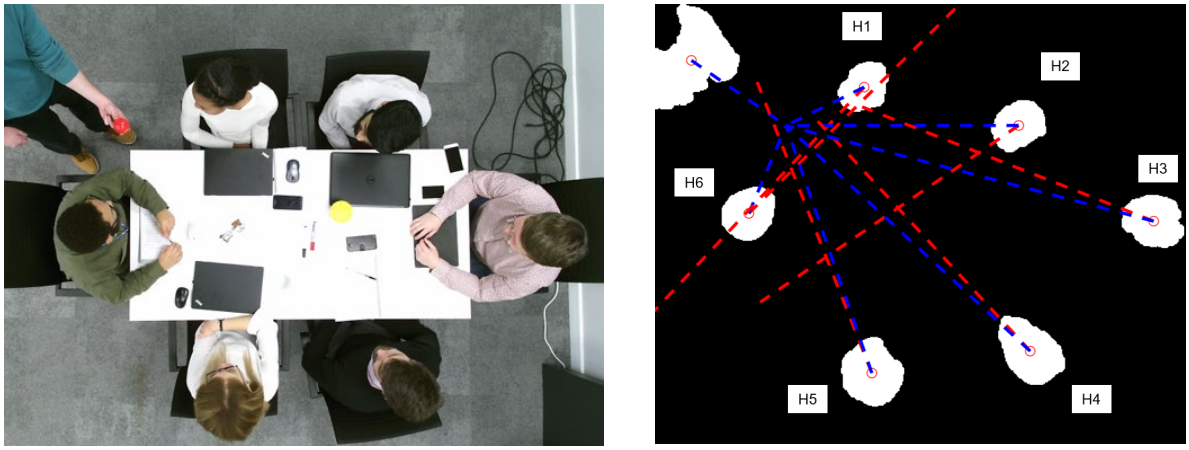


Figure 9. Plot showing the difference between manually measured head orientation (taken as ground truth) and the algorithmically estimated gaze direction angle (Experiment 1 data).

5.2.2 Evaluation of gaze estimation using the locations of the red ball

Fig. 10 shows a single frame comparing manual head orientation and computed gaze directions using the known location of the red ball. Head 2 (H2) highlights the impact on gaze direction when the participant is made to rotate their head beyond the comfortable limit of the range of movement to view an object out of their natural field of view (Stahl, 1999), and is likely they move their eyeballs to make up the remaining rotation.



(a) (b)

Figure 10. a) RGB image of scene, red ball position (OTL) and top left of the image is the person holding the ball and is not part of the experiment. b) Comparison of manual measurement of head orientation estimated from each head centre to the known location of the red ball (blue dashed lines) labelled OTL in Fig. 5b and the gaze direction estimated by the algorithm (red dashed lines).

Table 4 compares the computed gaze direction for each head with the orientation of the line joining the head centroid to the known locations of the red ball (as shown in Fig. 5b) over 140 frames, as the ball is moved to different locations around the table. The difference in orientation angle (degrees) between the two measurement methods is calculated for each frame and each head. The average across all pairs of angles was used to determine the accuracy of the gaze direction to the gaze fixation point. The results with the lowest angle difference, therefore the closest to the actual gaze direction, is for ball location ICN (inside centre -highlighted in green in Table.4), where the point of fixation, positioned in the centre of the table in front of all participants. In comparison, when the point of fixation, placed at position ITL (inside top left, highlighted in red in Table.4,) the results show high degrees of difference for some of the participants.

Ball Position	ID	H1	H2	H3	H4	H5	H6	Frames	Sec
Inside Top Left	ITL	11.5°(2.3)	27.5°(2.5)	2.7°(1.6)	3.3°(1.5)	1.2°(0.8)	20.1°(18.8)	28	6.0
Outside Top Left	OTL	14.9°(4.3)	35.1°(1.6)	7.3°(4.1)	4.2°(0.3)	1.8°(1.4)	22.1°(1.9)	16	4.0
Inside Top Right	ITR	12.7°(1.5)	4.7°(1.8)	7.1°(2.0)	9.2°(1.3)	13.0°(1.0)	9.8°(5.4)	18	4.0
Outside Top Right	OTR	10.9°(5.2)	6.9°(3.9)	4.0°(1.8)	11.0°(1.1)	20.0°(0.8)	9.1°(0.5)	23	5.5
Inside Bottom Right	IBR	3.3°(3.4)	4.8°(1.7)	14.1°(2.2)	8.6°(1.4)	35.7°(0.9)	10.3°(4.3)	6	2.0
Outside Bottom Right	OBR	3.6°(2.0)	11.7°(1.1)	3.6°(2.6)	19.1°(3.6)	46.9°(1.5)	5.3°(1.6)	8	2.0
Inside Bottom Left	IBL	23.4°(3.5)	16.6°(0.1)	11.0°(0.7)	16.8°(0.2)	3.3°(0.9)	16.7°(0.3)	2	0.5
Outside Bottom Left	OBL	23.4°(6.1)	9.0°(0.9)	14.8°(8.2)	22.0°(0.5)	21.9°(2.4)	34.9°(11.8)	7	2.0
Inside Centre	ICN	4.7°(2.0)	9.9°(6.5)	14.2°(7.3)	1.8°(1.3)	9.2°(0.7)	1.9°(1.1)	32	8.0
Mean		**10.4°(6.2)**	**15.1°(11.4)**	**7.9°(6.5)**	**7.4°(6.0)**	**12.9°(12.2)**	**12.5°(12.7)**		

Table 4. The mean angle in degrees, the difference calculated from the gaze direction algorithm and calculated angle from the centroid of the head to the known location of the object (ball) for each head and the standard deviation for each ball position and head (Experiment 1 data).

It is evident from these results that extending natural head movement without additional movement of the body impacts some participants, e.g. heads 1, 2 and 6 (H1, H2, H6) are very close to the point of fixation, or they are leaning and twisting their body to see. Head orientation for these participants ranges from a mean of 11.5° (H1) to 27.5° (H2) in comparison to the participants with natural head movement having an angle range of 1.2° (H5) to 3.3° (H4).

5.3 Focus of Attention (FOA) on people

The activity was captured over a period of 12mins 48 seconds (1447 frames) as the object was passed between the participants, starting with H3 then passing in an anti-clockwise order (H2, H1, H6, H5, H4) after which it rested on the table in front of H3. A total of 8,682 measurements resulted from the 1447 image frames.

Table 5 shows the focus of attention activity during experiment 3 when each participant either looked towards another person or the object. The results suggest that participant (H2) was observed for 21% of the period by other participants while participants (H5 & H6) attracted 2% and 3% respectively, of the observations from others. The object was observed 36% of the period. The results show that the proposed approach can be used to infer and support analysis of an individual's focus of attention.

	H1	H2	H3	H4	H5	H6	Object	Object %	Total Frames
H1	0	0	202	217	87	0	941	31%	1,447
H2	0	0	295	841	94	26	191	6%	1,447
H3	410	243	0	29	13	163	325	11%	1,183
H4	163	792	45	0	0	38	409	14%	1,447
H5	158	650	37	0	0	4	596	20%	1,445
H6	1	52	746	94	8	0	544	18%	1,445
Total	732	1,737	1,325	1,181	202	231	3,006		8,414
Observations %	9%	21%	16%	14%	2%	3%	36%		

Table 5. The number of times each head observed another head or the object a valid gaze was detected (a shown in Table 1). The head observations % is the proportion of times the participant (head) was looked at by the other participants (heads). The object % is the proportion of times the object was looked at by the specific participant (head).

Because an individual's gaze provides insight into their focus of interest, it can be inferred from the Observations row in Table 5 that participant (H2) was contributing (talking and in discussion) the most. Conversely, participants (H5 and H6) were less active contributors, drawing less attention from the other participants. The results indicate that the proposed approach can be used to infer and support analysis of an individual's focus of attention on people.

5.4 Common Fixation Point (CFP) on the object

Fig. 11 shows the first 500 frames for the gaze direction of head 1 (blue line) as the object passed between the participants. The red line plots the relative orientation of the object location (A1-A8) from H1's centroid. The plot indicates head 1's focus of attention closely followed the object in these early stages of the experiment (A1-A4) but is less attentive to the object as it moved around the table (A5-A8). Table 5 shows H1 is the most attentive observer of the object, representing 31% of the total 3,006 observations of the object.

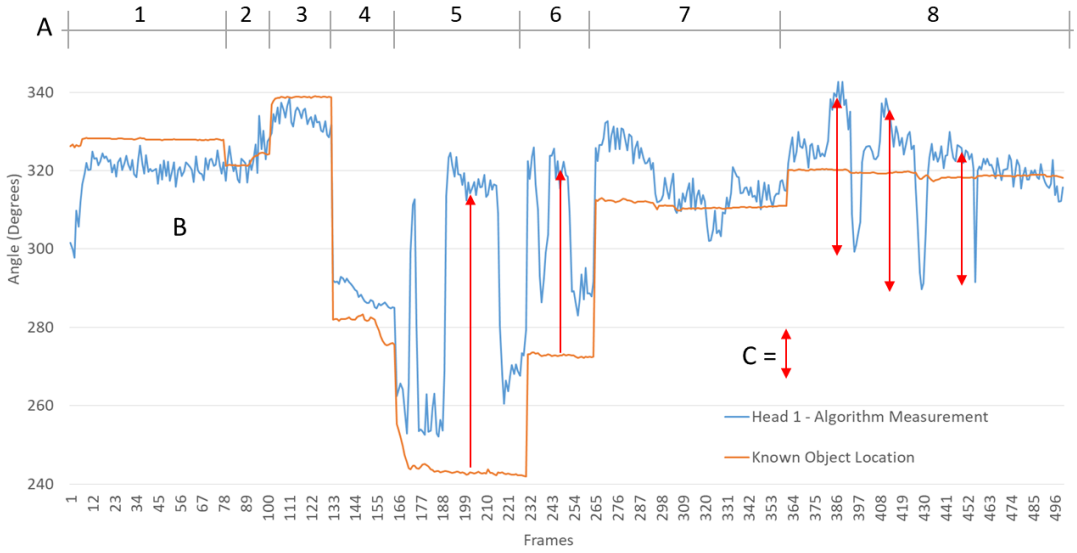


Figure 11. Plot combining head 1 gaze direction angles (blue line) with the angle from the centroid of H1 to the known location of the object (orange line). "A" section 1 to 8 indicate the change of location of the object and C (red arrows) indicate changes in head 1 movement away from the known object location. The slight changes seen in the known location line is the changes in the location of the head (taken from the algorithm output) to the known location of the object.

When the object moved to location A5 and A6, head 1 begins to move their gaze away from the object indicated by the red arrows C. This movement is completely away from the object and suggests the inferred focus of head 1 is on one of the other participants in the collaboration. In section A8 head 1 shows a different activity of inferred focus with head 1 movement away from and back to the object three times before settling back on the object, potentially suggesting a discussion on a particular element of the object with one of the participants.

Fig. 12 below shows a sequence for frames 265 to 500 of the data capture with the angle difference from the known object location and the inferred FOA for all six heads. The sequence of frames corresponds to sections A7 and A8 in Fig. 11 for all six heads. The plot illustrates the changes in the participant's focus, with the natural discussion taking place as the object was passed around the group. Head 2 and 3 show the most changes of the inferred focus of attention with a 40°- 60° change in angles from the object in frames 358 to 394 of the sequence.

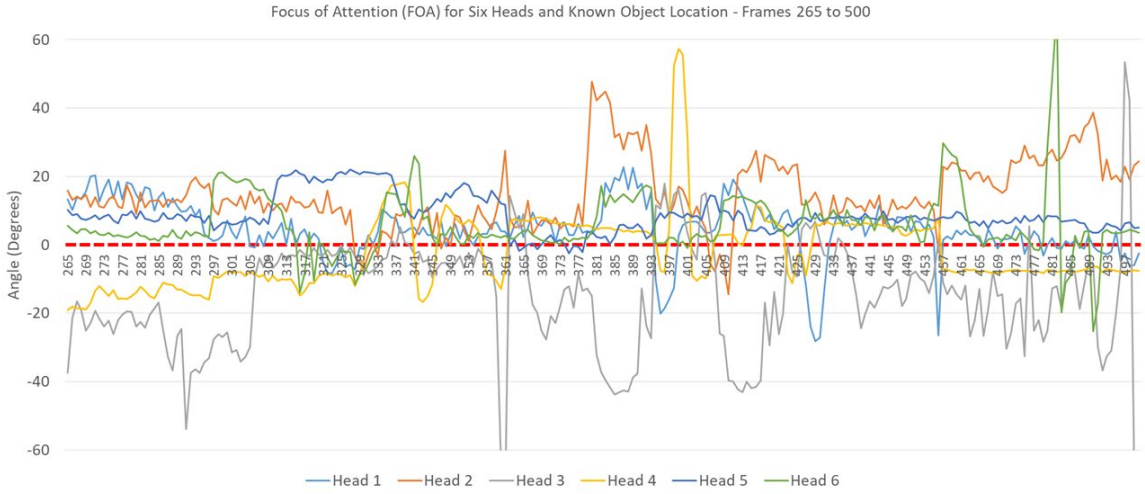


Figure 12. Plot combining the FOA angles for all six heads with the known object location set at zero. The difference in angle for each head from the known object location is plotted for frames 265 to 500 (corresponding to sections 7 and 8 in Fig. 11) to illustrate the changes in the inferred focus of attention for the six participants.

Fig. 13 plots the Euclidean distance between the inferred CFP computed from the intersection of gaze directions (i.e. the green star in Fig. 4b) and the known (approximate) location of the object (the blue star in Fig. 4a) in pixels over the first 500 frames of the sequence. Each pixel at the surface of the tabletop is 5.6mm in real world distance. The colours denote the change of location as the object passed around the participants (as labelled in Fig.6). Location A1-A2 covers the period when the object is initially introduced to the meeting (by H3) and placed on the table, and the attention of the participants is drawn to the object. During this change in group focus the distance drops from approximately 100 pixels (560mm) to 30 pixels (168mm). The groups attention continues as H2 examines the object (location A3). At locations A4-A6 the plot indicates no focus of attention on the object as H2, H1 and H6 hold it (a range of approximately 100 pixels (560mm) to 210 pixels (1,176mm). As the object moves to the final participant and is placed back onto the table (A7, A8) attention moves back to the object.

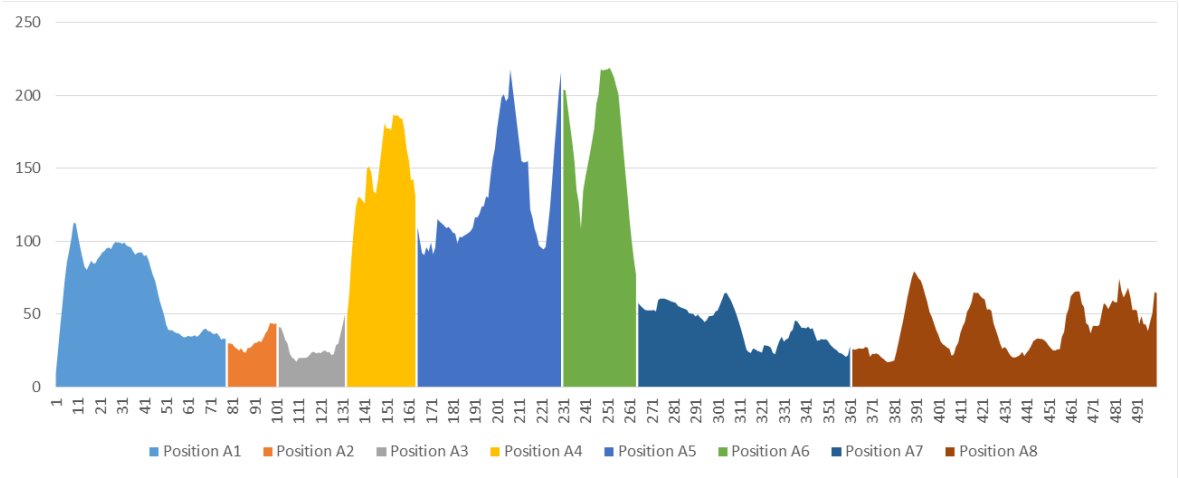


Figure 13. The plot shows the distance (in pixels) between the known location of the object (blue star Fig. 4a) and the CFP (green star Fig. 4b) of the six participants for the first 500 frames. One pixel is equivalent to 5.6mm in real world distance at the surface of the tabletop.

6. Conclusions

The results demonstrate the capability of a single depth sensor mounted overhead to monitor the participants in a constrained scenario (i.e. around a meeting table). This approach is unobtrusive, minimising the Hawthorne effect, simplifies data collection and avoids the need for multiple sensors. This makes it a viable solution to be used in general office meeting environments. As a consequence of the lack of comparative data from the literature (for which results are largely presented in a qualitative way) a detailed reporting of quantitative results of the effectiveness of our methods has been presented.

The novel combination of an overhead depth sensor and the MSER algorithm is shown to be very effective at detecting heads, with a detection performance of 99.94%, with only 5 segmentation failures. The ellipse fit to the detected head region resulted in a valid gaze estimation performance

of 96.9% of the head detections. The errors in estimating the head orientation from the best-fitting ellipse are associated with segmentation failures described in the previous paragraph, where the fitted ellipse can be distorted enough to rotate the estimated gaze through 90 degrees.

Evaluation of the head orientation measurements against manual estimates demonstrates errors ranging from 3.8° to 12.8°. It was found to be beneficial to incorporate this uncertainty into determining the Common Fixation Point (CFP) of the group of participants, allowing a Gaussian-weighted computation of the most likely location.

It is evident in the results that there are limiting factors to using horizontal head orientation (yaw) to estimate gaze direction. The limitations of using head orientation to estimate the gaze in situations where it exceeds the natural (or comfortable) extent of head rotation is observable in the results. In this case, the additional rotation would be made by the eye, or a person might otherwise rotate their body (or their chair), e.g. twisting to see. In addition, gaze direction does not explicitly identify the subject of the gaze, merely that the person is looking towards a particular subject. Hence, during the object discussion, when the object was in front of a person there would be no way to determine if an individual's FOA was on the object or the person if the viewer were directly opposite.

The proposed methodology of data capture and measurement has been shown to support the analysis of collaborative activities by estimating the Focus of Attention (FOA) of participants on other people and objects in an unobtrusive way. The experimental results presented a preliminary analysis of the behaviour of the participants and in particular the extent of their participation with each other and the point of discussion person/object within the collaboration environment. This technique is more suited to the detection of group attention, to identify group focus on a speaker or an object of discussion.

Further research will be undertaken to support the measurement of specific behaviours, such as loss of focus or inattention and distraction, and to understand the behaviour of people as they interact with each other. A goal will be to establish "attention profiles" of participants; identify the dominant and more passive participants in the interaction.

References

Ba, S.O. & Odobez, J.-M., 2006. Head pose tracking and focus of attention recognition algorithms in meeting rooms. In International Evaluation Workshop on Classification of Events, Activities and Relationships. pp. 345–357.

Bhattacharya, I. et al., 2018. A Multimodal-Sensor-Enabled Room for Unobtrusive Group Meeting Analysis. In Proceedings of the 2018 on International Conference on Multimodal Interaction. pp. 347–355.

Bhattacharya, I., Foley, M., Ku, C., Zhang, N., Zhang, T., Mine, C., . . . Welles, B. (2019). The unobtrusive group interaction (UGI) corpus. Proceedings of the 10th ACM Multimedia Systems Conference, (pp. 249-254).

Chong, E. et al., 2018. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In Proceedings of the European conference on computer vision (ECCV). pp. 383–398.

Cohen, I., Garg, A. & Huang, T.S., 2000. Vision-based overhead view person recognition. In Pattern Recognition, 2000. Proceedings. 15th International Conference on. pp. 1119–1124.

Daar, M. & Wilson, H.R., 2012. The face viewpoint aftereffect: Adapting to full faces, head outlines, and features. Vision Research, 53(1), pp.54–59.

Fischer, T., Jin Chang, H. & Demiris, Y., 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European Conference on Computer Vision (ECCV). pp. 334–352.
Ghiass, R.S. & Arandjelovic, O., 2016. Highly accurate gaze estimation using a consumer RGB-D sensor. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. pp. 3368–3374.
Hadjakos, A., 2012. Pianist motion capture with the Kinect depth camera. In Proceedings of the Sound and Music Computing Conference. pp. 303–310
Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions. PloS one, 10(8).
Hu, G. et al., 2014. Dt-dt: top-down human activity analysis for interactive surface applications. In Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces. pp. 167–176.
Kaminski, J.Y., Teicher, J. & Shavit, A., 2006. Head orientation and gaze detection from a single image. In International conference on computer vision theory and applications.
Kluttz, N.L. et al., 2009. The effect of head turn on the perception of gaze. Vision Research, 49(15), pp.1979–1993.
Mareschal, I., Calder, A.J. & Clifford, C.W.G., 2013. Humans have an expectation that gaze is directed toward them. Current biology : CB, 23(8), pp.717–21.
Masse, B., Ba, S. & Horaud, R., 2017. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. IEEE transactions on pattern analysis and machine intelligence.
Matas, J. et al., 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and vision computing, 22(10), pp.761–767.
Reddy, V.K., 2016. Estimation of Visual Focus of Attention from Head Orientations in a Single Top-View Image. pp. 43-47
Roethlisberger F. J. and Dickson W. J., Management and the Worker (Cambridge: Harvard University Press, 1939).
Stahl, J.S., 1999. Amplitude of human head movements associated with horizontal saccades. Experimental brain research, 126(1), pp.41–54.
Stiefelhagen, R., 2002. Tracking focus of attention in meetings. In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces. pp. 273-280.
Tian, Y. et al., 2003. Absolute head pose estimation from overhead wide-angle cameras. In Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on. pp. 92–99.
Voit, M. & Stiefelhagen, R., 2008. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In Proceedings of the 10th international conference on Multimodal interfaces. pp. 173–180.
Voit, M. & Stiefelhagen, R., 2010. 3D user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. pp. 51-58.
Wilson, H.R. et al., 2000. Perception of head orientation. Vision Research, 40(5), pp.459–472.
Wu, C.-J., Houben, S. & Marquardt, N., 2017. Eaglesense: Tracking people and devices in interactive spaces using real-time top-view depth-sensing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 3929–3942.