Head Pose Estimation is an Inadequate Replacement for Eye Gaze in Child-Robot Interaction

James Kennedy, Paul Baxter, Tony Belpaeme
Centre for Robotics and Neural Systems
Cognition Institute
Plymouth University, U.K.
{james.kennedy, paul.baxter, tony.belpaeme}@plymouth.ac.uk

ABSTRACT

Gaze analysis of human-robot interactions can reveal much about the dynamics of the interaction and be a useful step in establishing levels of engagement and attention. Currently, much of this work has to be conducted manually through post-hoc video coding due to current limitations in non-invasive, real-time gaze tracking solutions. This paper assesses whether real-time head pose estimation from an RGB-D camera may be used in place of manual post-hoc coding of gaze direction. Using data collected from an experiment 'in the wild', it is found that the proposed RGB-D based pose estimation method is neither accurate nor consistent enough to provide a reliable measure of gaze within human-robot interactions.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems

Keywords

Human-Robot Interaction; Child-Robot Interaction; Head Pose Estimation; RGB-D Camera; Video Coding

1. INTRODUCTION

Gaze analysis is used for many different applications in human-robot interaction (HRI). For example, gaze can be used as a measure to form engagement estimations, or as a proxy for attribution of social agency over time [1]. Such analysis is often completed through manual post-hoc video coding due to the current lack of non-invasive real-time eye tracking systems available for real-world use. This is a labour intensive task, thus making it desirable to automate the process.

As part of a larger study, the aim here was to evaluate an RGB-D camera (the Microsoft Kinect for Xbox 360, with SDK v1.7) for use as a means of automatically detecting whether a child is looking at a robot. Head pose estimation was used instead of eye gaze detection as it is more robust to lighting conditions [2]. It was hypothesised that in a relatively constrained physical interaction scenario, such head pose data would be strongly and consistently correlated with the actual gaze of the child (as measured through manual video coding), thereby providing a real-time estimation of gaze (thus facilitating means of influencing the interaction), and reducing the resource-intensive nature of data collection.

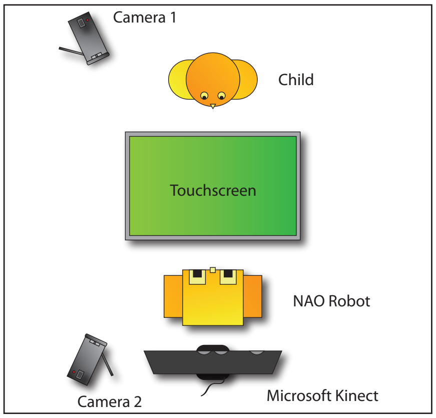

Figure 1: Schematic layout of the experimental setup. A seated child and an Aldebaran NAO interact over a touchscreen. A Microsoft Kinect is placed above and behind the robot, over 1 metre away from the child. A video camera records each interactant. Figure not to scale.

2. METHODOLOGY

Twenty interactions were filmed between children and the robot; all children had permission to take part and be filmed for use in the study. The children were all approximately 8 years old and from a primary school in the United Kingdom; 11 girls and 9 boys participated.

The robot was acting as a tutor for the children in educational interactions centered around a large touchscreen (Figure 1). The use of the touchscreen means that the movement of the child is relatively minimal and that their position reliably falls within the field of view of the cameras. The position of the robot head with respect to the RGB-D sensor is fixed and known. The robot provides lessons and feedback on touchscreen moves that the child made, as in [3]. An RGB-D camera was positioned slightly above and behind the robot, on a fixed mount, so that the robot did not obscure the image captured. The experiment was conducted in two

HRI'15 Extended Abstracts, March 2–5, 2015, Portland, OR, USA.
ACM 978-1-4503-3318-4/15/03.
http://dx.doi.org/10.1145/2701973.2701988.

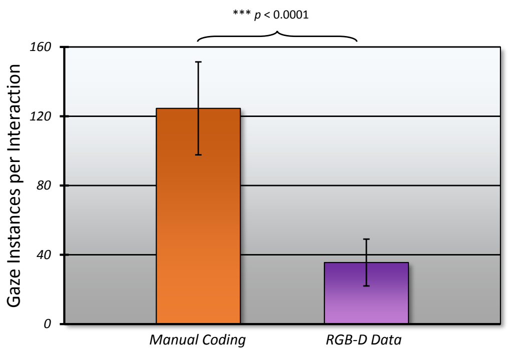

Figure 2: Average instances of gaze per interaction, comparing RGB-D data against manually coded data. * indicates significance at the p<0.001 level. Error bars show 95% confidence interval.**

different environments: one with ample natural light and one with minimal natural light.

The RGB-D camera was performing face tracking; when a face was detected, the head pose was estimated. The intersection of the estimated head direction vector and the robot frontal plane was calculated, and checked to see where the intersection point fell with respect to the robot head (this is known, given the fixed spatial relationship of the RGB-D camera and the robot head). Accuracy of the estimated gaze to robot was indicated by a confidence value ([0.0,1.0]) corresponding to the distance of the intersection point from the centre of the robot head in the frontal plane: a gaze event was logged if this value was greater than a threshold of 0.7 (corresponding to a distance of 12cm).

3. ANALYSIS AND DISCUSSION

The interactions were manually coded post-hoc by one coder, and validated by a second coder. 20% of the videos were second coded with an average Cohen's Kappa of 0.80, signifying substantial agreement [4]. For the RGB-D data to be useful as a replacement to manual video coding, it needs to be both accurate and consistent. The mean agreement between the RGB-D data and the hand-coded video data was a Cohen's Kappa value of 0.19 (95% CI [0.10,0.29]). This signifies 'slight' agreement [4] and suggests poor accuracy given the disparity with the human coded data.

Furthermore, a lack of accuracy is indicated when examining the total number of detected gaze events (Figure 2). The RGB-D data misses a significant number per interaction (M=35.6, 95% CI [22.0,49.1]) when compared to the manually coded data (M=124.6, 95% CI [97.7,151.4]); $t(19)$=6.872, p<0.0001. With the lowest Kappa value signifying less than chance, and the highest signifying 'substantial' agreement, it does not appear to be consistent across the interactions either. The highest Kappa value achieved for one of the twenty interactions was 0.65; Figure 3 shows the continuous RGB-D gaze confidence estimation against the manually coded value for this interaction.

Lighting conditions do not appear to be a factor as there were no significant differences between Kappa values for the minimally lit environment (M=0.118, 95% CI [0.00,0.24]) and the environment with ample natural light (M=0.25, 95% CI [0.10,0.39]); $t(18)$=1.421, p=0.172.

There are a number of ways in which the approach here could be developed to provide more accurate or reliable data. A general observation was that the children would often look towards the robot whilst their head pose was still oriented towards the touchscreen, which is reasonable based on human behaviour literature [5]. This would suggest that measuring eye gaze directly, rather than estimating gaze from head pose, may be necessary. In our setup, using an eye tracker positioned on the screen instead of behind the robot may be more effective.

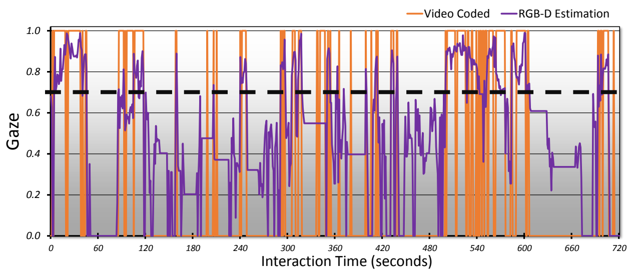

Figure 3: RGB-D gaze estimation confidence level compared to the binary manually coded values. The dashed line shows the threshold at which the confidence level is considered to constitute a gaze.

However, this does not mean that head pose data should be completely dismissed, since it has potential use as a means of generating basic reciprocal gaze behaviours, or other socially responsive behaviours, for the robot. Although a 'best case' from the data here, Figure 3 shows that there are many instances where a coarse behavioural response would likely be perceived as adequate by a human interactant.

Nonetheless, the results have shown that the head gaze data from the RGB-D camera is neither accurate, nor consistent. This suggests that it would not be suitable as a measure for HRI studies to replace manual coding.

4. ACKNOWLEDGMENTS

This work is partially funded by the EU FP7 ALIZ-E project (grant 248116), the EU FP7 DREAM project (grant 611391) and the School of Computing and Maths, Plymouth University, U.K.

5. REFERENCES

[1] P. Baxter et al. Tracking gaze over time in HRI as a proxy for engagement and attribution of social agency. In *HRI'14*, pages 126–127, 2014.

[2] G. Fanelli et al. Real time 3D head pose estimation: Recent achievements and future challenges. In *ISCCSP'12*, pages 1–4. 2012.

[3] J. Kennedy et al. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *HRI'15*, in press.

[4] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[5] M. Tomasello et al. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution*, 52(3):314–320, 2007.