Human-Machine Evaluation for Improving Gaze Estimation for Psychiatric, Child Populations and In-the-Wild Videos

Anonymous Author(s)

Affiliation

Address

email

Abstract

Human annotations are commonly used to improve the performance of machine learning (ML) models. However, previous work has shown that human annotations of gaze lack inter-rater reliability. If nothing else, machines are consistent. Thus, we propose a method for improving agreement between human raters using machine ratings to highlight how and why human raters differ. Our analysis furthermore provides an evaluation of state-of-the-art ML gaze estimation on in-the-wild videos of youth completing a psychiatric evaluation. We investigate parameters like the region of interest of the gaze of youth during diagnostic interviews, and gaze scores versus age. We find that ML methods can reveal differences between human rater scores and that both human raters’ and ML gaze estimates are lower for children aged 8-10 years than older youth.

1 Introduction

Eye movements and gaze are rich sources of information and provide important information about attention, engagement, and interaction between people [1]. Gaze analysis has important applications in psychiatry and psychology. The inability to make eye contact or maintain modulated eye contact can be indicative of mental disorders such as social anxiety, schizophrenia and autism [2, 3, 4, 5]. Humans can estimate gaze of another person with guidance from a behavioral coding manual, and can assign a global score that provides an average estimate of frequency, duration and intensity of gaze for a 3-minute video clip [6]. Humans can also give micro-codes for gaze, in which concrete behaviors are assessed, e.g., number of fixations and fixation duration, in a narrow time span [7]. Global scores have the advantage of being less time-consuming and tedious, whereas micro-codes may be more accurate and objective. Bias among human observations is a well-documented problem within behavioral research, with researchers in some fields avoiding the issue [8]. Efforts to reduce bias among raters include blinding raters and obtaining high inter-rater reliability [8]. Despite intensive training and practice, humans struggle to reach agreement on behavioral codes leading to high inter-rater variability in annotations [9].

Machine-based gaze tracking (a.k.a eye tracking) is a powerful tool for improving consistency, precision and speed within human behavior and psychiatric assessment [10, 11, 12, 13, 14]. Machine-based gaze tracking is ubiquitous and have become fairly reliable. However, the domain has mostly been explored for prospective applications, wherein dedicated equipment is used and tasks are designed to collect data that will be used to analyse gaze [15]. Work on in-the-wild clinical videos remains challenging, in terms of low-resolution videos, high variance and noise in recording conditions, limited training data and a requirement of person-independent (calibration-free) gaze estimation [16, 9]. In [17], the authors propose a tool to support the video coding of social attention in children with autism and in [18], the authors propose a method to extract interpersonal gaze from conventional

videos with an ambiguous physical layout. However, the challenges of gaze estimation for in-the-wild videos remain.

In this work, we take a step towards addressing these challenges by investigating the reasons for the divergence between expert annotations on gaze assessment of conventionally recorded videos of children from clinical populations. To the best of our knowledge, this is among the first works addressing this problem. Specifically, we study the influence of the region of interest (ROI) on the assessments made by experts and model gaze decisions. The influence of age and gender and their impact on the annotations and model output are also analyzed. The contributions of this paper are 1. an analysis of the factors that effect expert ratings. 2. a method to improve human rater agreement using feedback from a machine rater.

2 Data and Methodology

The methodology follows the same structure in terms of data and human estimation as in [9], and a brief description is provided in Appendix A.

![Figure 1: Face locations and ROI](image)

OpenFace [19, 20] is used with default parameters for gaze and head pose tracking. The eye gaze ($G_x$ and $G_y$) and head pose (pitch $P_{Hz}$, yaw $P_{Hy}$, and roll $P_{Hz}$) angles are given in world coordinates. These angles are in radians relative to camera position. The data set is challenging since the camera position varies, as shown in Fig. 1 (a), and the interviewer might cause occlusion when tracking the youth [18]. For quantifying gaze [9], the point of regard or the region of interest (ROI) is defined as a rectangle. ROI is bordered by the minimum and maximum values of the $G_x$ and $G_y$ that were extracted when the youth’s gaze was fixated on the eye region of the interviewer. The gaze score per video or Percent Inside (PI) is the percentage of gaze estimates falling within the ROI. PI is actually the percentage of time the youth looks at the interviewer. ROI is depicted in Figure 1 (b), where green points represent gaze estimates initially falling inside the ROI, and grey points are the gaze angles from the full video clip. Let $L_x$ and $L_y$ denote the length of the rectangle and x\%, y\% represent the increase in percentages over each axis. When evaluating gaze, the percent agreement (PA) [21] was used to measure the inter-rater reliability of gaze estimates:

$$\text{Percent agreement} (%) = \frac{\text{number agreements}}{\text{total number items}} \times 100$$

Agreement between raters was defined as a difference between two scores for the same video equal to or less than 1. Let total number items be the total number of videos. ML gaze percentages are transformed into scores with values ranging from 1 to 5 with half-points for the comparison with the raters. The percent exact match (PM) is computed in a similar fashion as PA, where the number agreements is replaced with number exact matches, i.e. the scores are the same.
3 Analysis and results

The percent agreement (PA) between Rater 1 and 2 was 64% which is considered low [9]. The ML approach with the ROI value set to zero compared to the human raters was also low: ML versus Rater 1 was 38% while ML versus Rater 2 was 52%. The ROI values have been increased and the effects on PA have been examined.

Figure 2 (a) illustrates that PA increased with ROI to a maximum level of 66% for Rater 1, while for Rater 2, PA continuously decreased as ROI increased. Figure 2 (b) shows that the maximum PM was obtained at around 150% for both raters. These results indicate that the ML approach with an increased ROI performs better for Rater 1. Figure 2 (c) illustrates that the highest value for PAs are obtained when PI is at 60%. This means that the enlarged ROI increased the fixation duration. Rater 1 has a smaller PI to start with 20% while Rater 2 has 25%.

The comparison between the two raters in Figure 2 indicates that the two raters are clearly rating differently when compared to the ML approach. A two sample t-test (Welch’s approach [22]) was used to test if the scores given by the raters are statistically different in terms of mean at a level alpha of 0.05. A p-value of 0.0006 was obtained indicating that there was a significant difference between the mean scores of the two raters.

We investigated the factors that may influence the difference between the raters. The gaze and head pose angles provided by OpenFace varies depending on the camera position. We extracted and computed the variance of the angles per video as it does not depend on the camera location. Furthermore, age, rater (score_rater) and the fixation duration (PI) were also examined. With the aforementioned features, a Principal Component Analysis (PCA) [23] was conducted to explore the variability observed in the video data.

Figure 3 illustrates the PCA biplot, where the first two components PC1 and PC2 explain 57% of the variation. As depicted, a visual separation is observed between the two raters. This is expected after conducting the t-test and including the raters’ scores. Moreover, lower scores seem to be concentrated at Rater 2. The loading’s direction indicates that Rater 1 used pitch and roll to estimate gaze while Rater 2 used yaw and gaze angles.

Removing the raters’ scores results in a PCA biplot depicted in Figure 4 (a), with 64% explained variance by the first two components. The separation among the groups disappeared visually.

A multivariate analysis of variance (MANOVA with Pillai’s trace criterion [24]) was conducted to check this difference statistically, assuming that the groups are multivariate normally distributed. At a significant level of alpha of 0.05, the groups are not different from each other since the computed p-value is 0.09. Thus, we do not expect the differences between raters to be caused by population differences in the annotated videos.
Furthermore, in Figure 4 (a) youth gender is also displayed and it can be visually denoted that there
is a male grouping with a visible spread. In Figure 4 (b), the mean scores versus the ages have been
computed with a ROI value of zero for the ML approach. As depicted, the scores are lower for
younger youth and higher when age increases. Rater 1 gives higher scores for older youth compared
to Rater 2. Younger youth tend to move more and maintain less gaze and therefore, the scores tend to
be lower. This also suggests that gaze tracking is more difficult for younger youth.

4 Discussion and conclusion

We studied the influence of the region of interest (ROI) on the expert and model gaze decisions.
It was observed that one rater’s agreement with the model increased as ROI increased up to 150%
whereas the other rater decreased in agreement with the model. The evaluation of the rater scores in
comparison to the ML model gives a clear indication of the difference in gaze estimates as a function
of the ROI. We also observed a difference in the agreement as a function of the percentage of gaze
scores falling inside the ROI (the PI) with a trade-off between the raters’ agreement to the model
at around 28% PI. According to the coding manual, consistent eye contact should result in higher
gaze scores and therefore the ROI is essential. The proposed method can be employed to improve
human rater agreement using feedback from a machine rater. Our method gives further feedback as
we analyse and present the factors that affect expert ratings, like ROI, pitch, roll, yaw (head pose) and
gaze angles. Based on the machine feedback, we observed that the gaze estimates of human raters are
influenced not only by the size of the ROI but also to a varying degree by the head pose.

The ROI for the human raters was larger than just the eye region of the interviewer. These PA values
indicate that a model with the ROI value set at zero is too limited of a plain of fixation. A ROI value
of zero represents eye contact between the youth and interviewer. This information is a valuable
take-away for model decisions. Age and gender were also incorporated in the analysis to explore
their impact on the annotations and model output. In particular, we noted that both raters and model
scored lower for younger children (8-10 years of age). Younger children often have a shorter attention
span, which could be one factor influencing these scores. An alternative explanation is that the model
was predominately trained on adults.

Estimating gaze in the wild has many useful applications, particularly in psychiatric assessment and
behavioral coding. Gaze is especially important in the study of attention, engagement, and relation-
ships in a clinical psychiatric setting. The first step towards developing effective machine learning
models is to understand how humans make their decisions, whereby we can design better models
that replicate the human thought process. We believe that a feedback loop between human raters
and the models could improve the inference of information from in-the-wild videos. Furthermore,
the challenges with high variability within ground truth labels from multiple raters can be addressed
by analyzing different human raters and understanding their rating process. ML tools can thus help
evaluate human decision inconsistencies in order to achieve more consistent results, which in turn
may improve annotations and aid in the improved alignment of models to raters.
References


A Appendix

A.1 Video Data

The video data set comes from a case-control study [25] and comprises psychiatric diagnostic screening interviews of 25 female and 12 male youth (8-17 years with a peak at 15) with (n=25) and without obsessive compulsive disorder (OCD) (n=12). The Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS) is a semi-structured interview designed to set psychiatric diagnoses in children and adolescents (6-18 years old) [26] [27]. Youth with an OCD diagnosis received outpatient care after assessment. Youth with clinical or subclinical tics or hyperactivity were excluded from this study. The interviews were recorded with a Sony video camera placed in different rooms for different participants with the camera placed in various locations. All videos had the youth as the main object of focus with interviewers only occasionally appearing in the frame. To avoid bias, the video clips were extracted by the first author and not the raters. Videos include the middle 30 seconds of the depression or mania chapters of the interviews. Thus, the data set includes 2 videos from each youth resulting in 74 videos.

A.2 Human gaze estimation

Two mental health professionals and coauthors (Raters 1 and 2) were trained in using the global scores of the Coding Interactive Behavior (CIB) manual [6] for 3-minute videos. On a separate set of videos, they reached 89% agreement. Sixty videos were scored by either Rater 1 or 2, whereas 14 videos were scored by both raters to calculate rater agreement. Gaze of the youth was rated using the "child gaze” item in the adolescent version of the CIB manual. Significant differences in CIB scores of youth with and without a psychiatric disorder has been documented [7].

Gaze is rated on a scale from 1 to 5. Whole- and half-points can be given (e.g., 1.5 or 3.5). To receive a maximum score, the child must focus gaze on the interviewer or an object of joint attention (in this case the interview questions) and maintain eye contact with the interviewer consistently during the interaction. Looking away from the interviewer for most of the interaction should receive the lowest score. Due to the time-consuming process of the scoring, each rater scored 44 videos in a random order. From these 44 videos, 14 videos were scored separately by both raters. All the 74 videos were scored by either Rater 1 or 2. Raters scored a batch of 7 videos at a time (the last session had 9 videos). The raters met and discussed codes after each batch to avoid coder drift. It was attempted to blind raters to diagnostic status and diagnostic interview chapter, but this information is often indirectly available in the video. For many, meta information is also available since the raters also made a description.