Detecting Bids for Eye Contact Using a Wearable Camera

Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, James M. Rehg
Center for Behavioral Imaging, School of Interactive Computing, Georgia Institute of Technology

***Abstract*— We propose a system for detecting bids for eye contact directed from a child to an adult who is wearing a point-of-view camera. The camera captures an egocentric view of the child-adult interaction from the adult's perspective. We detect and analyze the child's face in the egocentric video in order to automatically identify moments in which the child is trying to make eye contact with the adult. We present a learning-based method that couples a pose-dependent appearance model with a temporal Conditional Random Field (CRF). We present encouraging findings from an experimental evaluation using a newly collected dataset of 12 children. Our method outperforms state-of-the-art approaches and enables measuring gaze behavior in naturalistic social interactions.**

I. INTRODUCTION

Our paper presents a novel method for detecting a child's bids for eye contact with an interactive partner[1]. Eye contact is a powerful social signal and plays a crucial role in regulating social interactions from the first months of life [17]. Before they are able to speak, infants and toddlers communicate with others using well-timed looks coordinated with gestures and vocalizations [26], [14]. Eye contact also plays a key role in joint attention, used to denote a class of behaviors in which children use gaze and gesture to spontaneously create or indicate a shared point of reference with another person [28]. Beyond its importance and relevance to the study of typical development, eye contact represents a key area of focus for those studying autism. Atypical patterns of gaze, eye contact, and joint attention have been identified as among the earliest indicators of autism in the first two years of life [15], [32], [16], and continue to characterize individuals with autism throughout childhood and adolescence [36], [19].

In spite of the developmental importance of gaze behavior in general, and especially in social interactions, there currently exist very few good methods for collecting large-scale measurements of these behaviors in the course of naturalistic interactions. A classical psychology setting involves multiple static cameras recording the scene from different view points. This setup is feasible in a laboratory setting, but it is not amenable to collecting data over a long period of time or across a variety of different environments. In addition, it is extremely challenging for human coders using only environment-mounted cameras to accurately and reliably determine whether a child is making eye contact or looking at other parts of the face. An alternative solution is to use eye tracking methods, which are accurate and can be applied to large numbers of subjects. However, this approach requires a child to either passively view content on a monitor screen or wear a portable gaze-tracker. Neither of these are completely suitable for naturalistic, face-to-face interactions. Even if we can mimic a social interaction using monitors [13], it is not clear that the gaze behavior in such a setting will reflect gaze behavior in the real world [11].

To address the limitations of previous approaches, we propose to measure gaze behavior in naturalistic social interactions by using a point-of-view (POV) camera worn by the social partner of the person whose gaze is of interest. We note that such an egocentric setting is a particularly good vehicle for eye contact detection, as it provides high quality visual data, e.g. consistent near-frontal view faces, higher resolution eye regions and less occlusion. Our approach capitalizes on the increased availability of wearable POV cameras such as Google Glass[2] and Pivothead[3]. We begin with instrumenting the child's social partner with a pair of glasses containing a high-definition outward-facing camera. The camera is placed close to the social partner's eyes and aligned with his or her point of view, and naturally captures the child's face and looks toward the partner's eyes during an interaction. We apply facial analysis to the egocentric video and design a learning based approach to determine when the child is looking toward the camera, as an approximation of the child's looking toward the partner's eyes. We show that these events cover most of the child's bids for eye contact.

Detecting bids for eye contact in such videos is a challenging problem. The appearance of the human eye is quite diverse. For example, eye regions under two different head poses can look similar yet be perceived as different gaze behaviors [22]. We address this issue by learning a pose-dependent appearance model for detecting bids for eye contact at each frame. These frame level results are further combined into a sequential Conditional Random Field (CRF) [21] to model the temporal events of bids for eye contact. Our method thus provides both a frame level and an event level estimate of the child's bids for eye contact. To evaluate our method, we collected a dataset of 12 sessions consisting of a semi-structured play interaction between an adult and a child. Annotations from 5 coders indicate a high annotation consistency with a very small probability of missing the bids in the egocentric video. Our method reaches a high accuracy against human annotations on the dataset. Our benchmark also demonstrates that our approach outperforms state-of-the-art gaze estimation methods. For the rest of the paper, we use the terms "bid for eye contact" and "eye contact" interchangeably if no ambiguity occurs. It is worth noting

[1] http://cbi.gatech.edu/eyecontact/

[2] https://www.google.com/glass/start/

[3] http://www.pivothead.com/

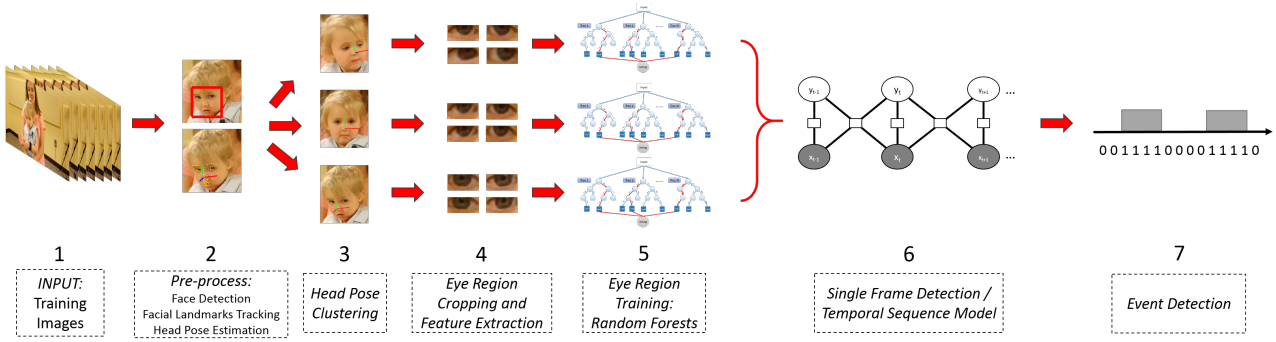


Fig. 1: Overview of our approach. **1**: Extract frames from an egocentric video; **2**: Perform face detection on each frame; track facial landmarks; estimate head pose. **3**∼**5** Training phase: Cluster faces based on head poses; crop eye regions based on facial landmarks and extract features from the cropped eye regions; train a binary classifier for each cluster independently; Testing phase: Find the appropriate classifier for eye region features via pose estimation. **6**: Aggregate frame level results and model temporal information via a linear-chain CRF. **7**: Detect events based on labels from CRF.

that the gaze direction of the camera wearer is not considered in our setting.

The contributions of this paper are: 1) the first dataset of the egocentric videos collected in the course of a naturalistic adult-child interaction; 2) a method for detecting bids for eye contact that uses pose-dependent appearance features and attains better performance than pose-independent approaches at the frame level in an egocentric setting; 3) an approach to detecting bids for eye contact at the event level by aggregating the frame level results and modeling the temporal structure via CRF. To the best of our knowledge, we are the first to detect bids for eye contact events in videos.

II. RELATED WORK

A. Appearance-Based Gaze Estimation

Much recent work on gaze estimation has focused on estimating gaze direction using static cameras [37], [39], [25], [40]. In particular, Smith et al. [37] use eye region images to train a classifier to detect if the user is looking at the camera. Sugano et al. [39] use a large amount of cross-subject training data to train a 3D gaze estimator, which is person- and head pose-independent. Schneider et al. [34] perform a manifold embedding for each person in the training dataset and learn a linear transformation to estimate gaze. All these approaches are trained and tested on datasets composed of images of adults' faces collected in a controlled and constrained setting. In contrast, our dataset consists of children's faces in egocentric videos during a naturalistic interaction with an adult.

The most relevant work is from [43]. Ye et al. [43] use wearable gaze-tracking glasses to detect adult-child eye contact by using commercially available gaze estimation software[4]. Our work differs from Ye et al. in three key areas: 1) we improve the performance of single frame eye contact detection by extracting appearance features from eye region images; 2) we consider the dependency between head pose and gaze direction and build a pose-dependent appearance model; 3) we embed the appearance model into a temporal CRF, which allows us to extract eye contact events in an egocentric video.

B. Egocentric Vision

There is a growing interest in using wearable cameras in computer vision, motivated by advances in hardware technology. Most work has focused on understanding the first person's actions/activities [38], [8], [18], [44]. Only a few works address social interactions [9], [33] or gaze behavior [24]. Fathi et al. [9] model the visual attention of people in a social setting via face detection and head pose estimation. Ryoo and Matthies [33] integrate global and local motion features to recognize the physical interactions between a human and a robot equipped with a camera. Li et al. [24] estimate the first person's gaze direction by leveraging the implicit cues in the camera wearer's head movement. None of these previous works address the modeling of eye contact during a social interaction.

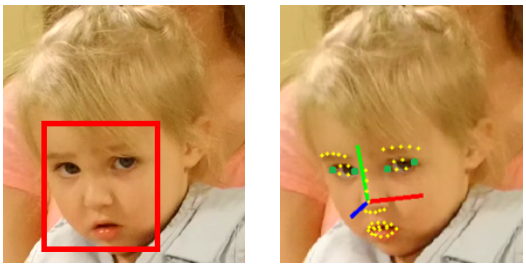
C. Eye-Tracking for Identifying Developmental Disorders

A large body of behavioral research indicates that individuals with diagnoses on the autism spectrum have atypical patterns of eye gaze and attention, particularly in the context of social interactions [5], [23], [35]. Eye-tracking studies using monitor-based technologies suggest that individuals with autism, both adults [20] as well as toddlers and preschool-age children [3], [16], show more fixations to the mouth and fewer fixations to the eyes when viewing scenes of dynamic social interactions as compared to typically developing and developmentally delayed individuals. Importantly, atypical patterns of social gaze may already be evident by 12 to 24 months of age in children who go on to receive a diagnosis of autism (e.g. [46], [41]).

D. Eye-Tracking for Interaction with Children

Several methods for gathering eye-tracking signals from infants in the course of naturalistic interactions have been

[4] http://www.omron.com/r_d/coretech/vision/okao.html



(a) OMRON (b) IntraFace

Fig. 2: Face Detection using OMRON OKAO Vision is shown in Fig. 2a. Facial landmarks tracking and head pose estimation using IntraFace are shown in Fig. 2b. The combination of OMRON and IntraFace provides accurate eye corner (green dots) localization and head pose estimation.

proposed [29], [12], [13], [45]. In particular, Noris et. al [29] present a wearable eye-tracking system for infants and compare the statistics of gaze behavior between typically developing children and children with autism in a dyadic interaction. Yu and Smith [45] ask parents and their children to wear a head-mounted eye tracking system during toy play to examine how they coordinate attention to objects held by the self vs. the social partner. Guo and Feng [13] measure child-parent joint attention during a storybook reading task by showing the same book on two different screens and simultaneously tracking the gaze of the parent and child with two separate eye trackers. However, these previous studies either relied on a specially designed wearable eye trackers [29], [12], [27], or limited the interaction to a computer screen [13]. Our method addresses the problem of bids for eye contact, and enables detection of eye contact in a naturalistic interaction *without instrumenting the child.*

III. APPROACH

Fig. 1 presents an overview of our method. We take an egocentric video as input and detect faces in each frame. Face detection is followed by facial landmark tracking and head pose estimation. Eye regions are cropped using the landmarks and features are extracted from cropped regions. We divide the faces into clusters based on their poses and train a binary classifier for each pose cluster independently. At the testing phase, the classifiers are applied at each frame and the results are aggregated. We further combine single frame results into a temporal CRF. We use Forward-Backward algorithm to perform inference in the CRF model. Our final output is a set of eye contact events in an egocentric video.

A. Pre-Processing

We use the OMRON OKAO Vision software[5] to detect the child's face in the egocentric video captured by a camera worn by the social partner. OMRON OKAO Vision is a commercial facial analysis software that includes face detection. For each frame, the software outputs a bounding box as the face location in the image (Fig. 2a). Using the bounding box, we further track facial landmarks and estimate head pose of the detected face using a publicly available face alignment system, IntraFace[6] [42]. IntraFace extends the cascaded pose regression [6] with a Supervised Descent Method (SDM) for tracking facial landmarks. We also estimate head pose using the tracked landmarks. The head pose is composed of three degrees of freedom — yaw, pitch and roll. Fig. 2b shows a visualization of tracked facial landmarks and estimated head pose. The green dots in Fig. 2b indicate the corners of left and right eyes. The combination of OMRON face detector and IntraFace face alignment tool offers best results for face detection and facial landmarks (including eye corners) tracking. Yet, both OMRON and IntraFace, designed for a static camera, can only find 83.02% of frames that are identified as eye contact by human annotators due to motion blur, occlusions, and other factors (see Sec. IV-B for more details). However, when a face is detected, the landmarks are highly accurate.

B. Frame Level Eye Contact Detection

We first present our method for detecting eye contact at each frame. Eye contact detection is implemented as a binary classification over every frame in the video. The label can be either eye contact or not. Our single frame model consists of two parts: clustering head poses and modeling pose-dependent appearance features. We detail each step here.

1) Head Pose Clustering: We estimate the head pose h_t at frame t using the landmarks from IntraFace by registering the landmarks into an average 3D face. The appearance of the human eye region varies significantly under different head poses, even during moments of eye contact from the same individual. Fig. 3 illustrates an example where the appearance of the eye regions stays the same, yet humans perceive different gaze behavior under different head poses. To address this issue, we propose a pose-dependent appearance model inspired by [40]. The idea is to model the appearance of eye regions separately for each head pose. We first cluster head poses using Gaussian Mixture Model (GMM) and train a classifier for each pose cluster using appearance features. We only cluster the head poses using pitch and yaw, as the rotation can be resolved by a simple affine transformation. The number of clusters is set to 3 for all our experiments.

2) Feature Extraction: Using facial landmarks from IntraFace, the eye regions (left and right) are cropped from each frame using eye corners. We align the tilted eye regions using affine transformation and resize the regions into a fixed resolution of 73×37. Affine transformation cancels the in-plane rotation and makes eye region rotation invariant. Fig. 3 shows the cropped eye regions. To generate more training samples, we randomly perturb the eye corner locations by an offset of 2 pixels. From the cropped eye regions, we extract appearance features A_t at frame t by concatenating the feature vectors from both eyes. We experimented with LBP [30],

[5] http://www.omron.com/r_d/coretech/vision/okao.html

[6] http://www.humansensing.cs.cmu.edu/intraface/

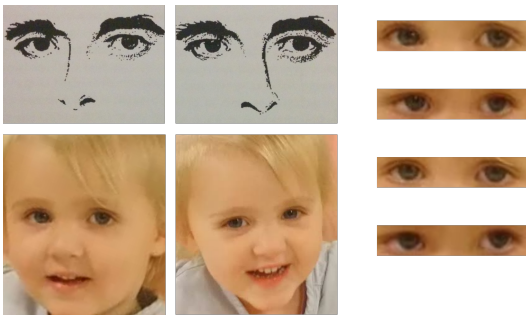
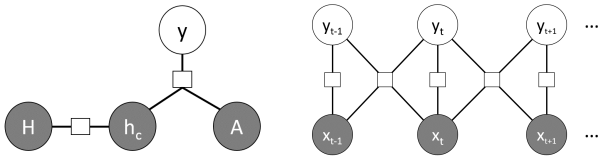


Fig. 3: Top left: A sketch from [22] demonstrates the dependence between head pose and gaze behavior. The two eye regions are very similar yet only the left one is perceived as eye contact. Bottom left: An example from our dataset. The two frames are both marked as eye contact yet the visual appearance looks quite different. Right: eye regions from our dataset. It is very difficult to tell if it is an eye contact given only the eye region.



(a) Factor Graph Representation of Single Frame Eye Contact Detection Model.

(b) Linear-Chain CRF Representation of Sequence Eye Contact Detection Model.

Fig. 4: Graphical model for eye contact detection. Fig. 4a: single frame model. y is the binary label of eye contact at each frame, A represents appearance features from eye regions, H is the head pose given by landmarks, and h_c is an indicator variable for head pose cluster c. Our model extracts pose-dependent appearance features for single frame eye contact detection. Fig. 4b: CRF for events. y_t is the binary label of eye contact at frame t, and x_t consists of both appearance feature and head pose at frame t. Our model smoothes the single frame results by a Potts model.

CNN feature [2], and HOG from [10], which is a modified version of the original Histogram of Oriented Gradients [4]. Results can be found in Sec. IV-B.

3) Single Frame Eye Contact Detection: We consider single frame eye contact detection to be a binary classification problem. We denote the input head pose as H_t and the appearance feature as A_t at frame t. We give a binary label y_t to frame t, where 1 indicates eye contact and 0 otherwise. The conditional probability $P(y_t|H_t,A_t)$ is modeled as

$$P(y_t|H_t,A_t) = \sum_{h_c=1}^{k} P(h_c|H)P(y_t|h_c,A), \tag{1}$$

where h_c is the indicator variable for c-th cluster of head pose in GMM. The key intuition is to consider each cluster as a sub-problem and average the results of the classifier from each cluster at the last step. h_c can also be considered as a latent variable where (1) is interpreted as a Bayesian inference. We choose Random Forests [1] as our classifier for $P(y_t|h_c,A)$ with 100 trees and 10 splits per node. Fig. 4a illustrates the factor graph of the model. We further consider the negative log-likelihood $u(x_t,y_t)$ of $P(y_t|H_t,A_t)$ as

$$u(x_t,y_t) = -\log P(y_t|x_t) = -\log \sum_{h_c=1}^{k} P(h_c|H)P(y_t|h_c,A), \tag{2}$$

where x_t represents both head pose and appearance feature at frame t, and $u(x_t,y_t)$ is used later as the unary potential in our temporal model — a linear-chain CRF.

We name our single frame eye contact model as **P**ose-dependent **E**gocentric **E**ye **C**ontact detection, or **PEEC**.

C. Event Level Eye Contact Detection

Our ultimate goal is to find onsets and offsets of eye contact events within an egocentric video. This requires us to move beyond single frame eye contact detection. We consider a simple linear-chain CRF as the temporal model. The CRF smoothes the single frame results and produces more accurate onsets and offsets of the eye contact events.

We expand our single frame model (PEEC) (Fig. 4a) to construct a temporal Conditional Random Field (CRF). The CRF represents the event detection as sequential labeling. Fig. 4b illustrates the linear-chain CRF of our sequence model. Denote n as the total number of frames. The model can be expressed as

$$\begin{aligned} P(\{y_t\}|\{x_t\}) &= \prod_{t=1}^{n} P(y_t|x_t) \prod_{t=1}^{n-1} P(y_t,y_{t+1}|x_t,x_{t+1}) \\ &= \frac{1}{Z}\exp(-E(y_1,y_2,...,y_n,x_1,x_2,x_n)), \end{aligned} \tag{3}$$

where $E(\{x_t\},\{y_t\})$ is the energy function and Z is the normalization factor. We further simplified the model by using a Markov assumption. Now we have

$$E(\{x_t\},\{y_t\}) = \sum_{t=1}^{n} u(x_t,y_t) + \sum_{t=1}^{n-1} v(y_t,y_{t+1},x_t,x_{t+1}). \tag{4}$$

$u(x_t,y_t)$ is given by the negative log likelihood in (2). $v(y_t,y_{t+1},x_t,x_{t+1})$ is the temporal pairwise potential. We use the standard Potts model for the pairwise term as

$$\begin{aligned} &v(y_t,y_{t+1},x_t,x_{t+1}) \\ &= \lambda I[y_t \neq y_{t+1}] \exp\left(-\frac{\|H_{t+1}-H_t\|_2^2}{\sigma_H^2} - \frac{\|A_{t+1}-A_t\|_2^2}{\sigma_A^2}\right). \end{aligned} \tag{5}$$

Our definition in (5) serves as a penalty function that activates when the labels of two adjacent frames are different. We use Euclidean distance (ℓ^2-norm) as a metric to measure the differences of the head pose H_t,H_{t+1} and the appearance feature A_t,A_{t+1} between two adjacent frames.

By minimizing the energy function in (4), we get the solution of CRF as sequential labels.

$$y_1^*,y_2^*,...,y_n^* = \underset{y_1,y_2,...,y_n}{\operatorname{argmin}} E(y_1,y_2,...,y_n,x_1,x_2,...,x_n). \tag{6}$$

We can obtain events of eye contact by merging a set of consecutive $1's$ over the results of CRF.

Learning: Since there are only three parameters, λ, σ_H

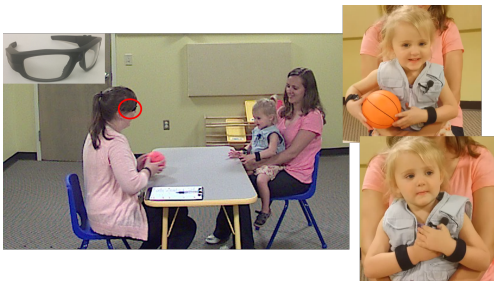


Fig. 5: Data Collection Setup. Left image: the adult (seated on the left) wears a pair of video-recording glasses and interacts with the child (seated on the right). Right images: two sample frames from egocentric videos.

and σ_A, in our model (5), we perform a grid search over a separate validation set. The search is limited within $[0.1, 10]$, which is equally spaced in log scale.

Inference: Given the unary and pairwise potentials, we can detect eye contact event by inferring the best label of every single frame. The linear-chain CRF allows us to use the efficient Forward-Backward algorithm for the inference task.

IV. Experiments and Results

We conduct a series of experiments and report the results in this section. We begin by the introducing a new dataset of dyadic social interaction between an adult and a child. We further report both frame level and event level results of our method on the dataset, in comparison to state-of-the-art methods. Our method outperforms all previous methods and attains high accuracy when compared against human annotations.

A. Data Collection and Annotation

We collected egocentric videos from 12 adult-child interactions. Child subjects ranged in age from 18 to 28 months, an age interval selected as it is relevant to early detection of developmental disorders such as autism. Each session consists of a 3-5 minute table-top play interaction between a child and an adult examiner, leading to over 85K frames in total. The child was seated across from the adult at a small table. The adult engages the child in various activities, following the same protocol as our earlier MMDB work[7]. See [31] for a detailed description of the protocol. All child subjects participated with consent from their parents. Videos were recorded by a pair of Pivothead glasses worn by the adult while interacting with the child. The resolution of the video is 1280×720 with 30 frames per second. Fig. 5 shows the setup of a data collection session. In this setting, we observe a high quality image of the child's face in the egocentric video recorded by the glasses.

Each session was independently scored by five different annotators to flag frame level onsets and offsets of each instance of the child bids for eye contact with the examiner. The annotators used ELAN, an open source video annotation tool[8], and viewed three synchronized video feeds from each session: (1) the egocentric video recorded by the Pivothead glasses worn by the adult; (2) an HD video recorded from a stationary camera placed behind and to the left of the examiner such that it captured the child's face and the back of the examiner's head, and (3) an HD video recorded from a stationary camera placed behind and to the left of the child such that it captured the examiner's face and the back of the child's head. The annotators thus relied on all three videos in making their determination whether the child looked into the examiner's eyes. The videos together with the annotations will be made available to the research community.

Session ID	Age (Month)	Unscorable (%)	Consistency
1	23	0.42	0.72
2	23	1.16	0.91
3	19	4.00	0.85
4	17	4.46	0.80
5	28	1.53	0.90
6	24	4.48	0.92
7	24	4.98	0.89
8	24	0.00	0.90
9	22	2.73	0.90
10	20	8.13	0.86
11	29	8.90	0.84
12	18	4.97	0.88
Average	22.48	3.82($\pm$2.82)	0.86($\pm$0.06)

TABLE I: Statistics of each session in our experiments. The third column (Unscorable) indicates the percentage of frames the child's eye contact cannot be determined based on the egocentric video (because the child's eyes are off camera), but can be determined using the videos from stationary cameras. The fourth column (Consistency) indicates the average Intersection over Union among different annotators.

Given multiple annotations of the same session, we measure the annotation consistency by comparing the Intersection, the frames where both annotators marked as eye contact, over Union, the frames where that any annotator marked as eye contact, between each pair of annotators (IoU). These pairwise consistencies are then averaged for each session and reported in the last column of Table I as "Consistency". The average IoU over all sessions is 86.42%, indicating a very high consistency among our annotators. This allows us to use a simple majority voting on every frame to obtain ground truth eye contact.

There are moments when the child's eyes are not visible in the egocentric video but a determination of the presence or absence of eye contact can still be made from the stationary cameras. Thus the annotators further flagged the frame level onset and offset of any period when the child's eyes were not visible in the egocentric video. We explore the impact of such events when evaluating the accuracy of the automated eye contact detection against ground truth annotation. We measure the average percentage of frames marked as eye contact while the child's eyes are missing in the egocentric video, denoted as "Unscorable" in Table I. On average, only 3.82% of the frames are missing. This small percentage

[7] http://cbi.gatech.edu/mmdb/

[8] http://tla.mpi.nl/tools/tla-tools/elan/

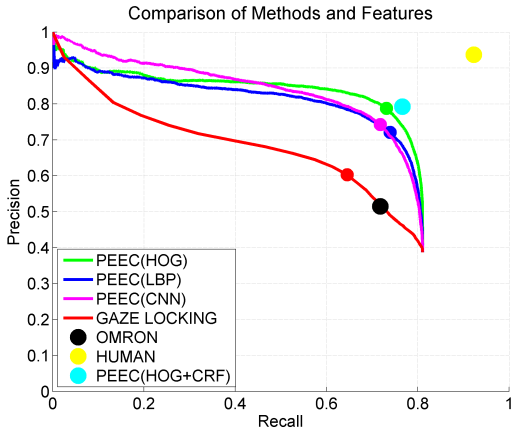


Fig. 6: Average PR curve of different methods and features. The dot on each PR curve indicates the threshold picked by the corresponding maximum F_1 score.

suggest that 1) egocentric video is able to capture almost all of the eye contact events in our setting; 2) we can safely exclude these frames from our benchmark, as they won't have significant impact on the results.

B. Single Frame Eye Contact Detection

Our first experiment benchmarks the performance of single frame eye contact detection. The detection is considered as a binary classification problem at every frame. As the frames marked as eye contact only occupy a small portion of the data, we use Precision-Recall (PR) curve and F_1 score as our evaluation criteria. More precisely, we have

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{7}$$

where $precision = \frac{tp}{tp+fp}$, $recall = \frac{tp}{tp+fn}$, tp is the number true positive samples, fp is number of false positive samples, and fn is the number of false negative samples. To fairly compare the performance across methods, we use leave-one-subject-out cross validation for all experiments. The results are thus reported by averaging all 12 folds.

We compare PEEC against two baseline state-of-the-art methods in [37] and [43]. We also test different features using our pipeline, including LBP, HOG and CNN features. The CNN feature we use is pre-trained on the ImageNet 2012 dataset, and then extracted from the $7th$ layer without fine-tuning [2]. All results are reported on our dataset. Fig. 6 shows the PR curve for our method with different features, in comparison to [37] (Gaze Locking), [43] (OMRON), and human consistency. Our method with CNN features performs slightly worse than our methods with HOG features. This is probably due to the relatively low resolution of eye region images and the lack of fine-tuning. Since HOG outperforms all other features, we choose HOG as the appearance feature for our PEEC model in all subsequent analyses.

We also perform significance tests between PEEC and two baseline methods. The tests are based on the F_1 scores of all 12 sessions using t-test. The p-value between PEEC and OMRON is 0.00020, and the p-value between PEEC and Gaze Locking is 0.00003; both p-values are less than 0.05.

(a) Face detection failed due to partial occlusion.

(b) Face detection failed due to face rotation.

Fig. 7: Our method is not able to detect eye contact when face detection is failed. These failure cases occupy 16.98% of total eye contact frames in our dataset.

	Precision	**Recall**	**F_1 Score**
OMRON [43]	0.5151	0.7179	0.5998
Gaze Locking [37]	0.6028	0.6454	0.6234
PEEC	**0.7929**	0.7268	0.7584
PEEC+CRF	0.7920	**0.7664**	**0.7790**

TABLE II: Performance Comparison among OMRON, Gaze Locking, PEEC, and PEEC+CRF.

We can further improve the PEEC frame level result by adding CRF. Table II shows the precision, recall and F_1 scores of two baseline methods, PEEC, and PEEC+CRF at their best thresholds. PEEC+CRF achieves the best recall and F_1 score.

As shown in Fig. 6, PEEC+CRF does not attain human-level performance. This is partly due to the reason that causes all PR curves in Fig. 6 to not reach $(1,0)$: in our dataset, 16.98% of the faces that are identified as eye contact by human annotators cannot be detected by our face detector. Some of these faces are partially occluded during the interaction, nonetheless, humans are still be able to identify eye contact from them (see Fig. 7a). Other failure cases are caused by extreme face rotation (see Fig. 7b), due to either the camera pose (the head pose of the camera wearer) or the head pose of the child. Detecting the child's face in egocentric videos is a major bottleneck of our method, and likely results from the fact that most current face detection methods were trained on images of adult faces captured by stationary cameras.

C. Event Level Eye Contact Detection

We take one step further to benchmark the event detection results. Similar to object detection in computer vision, we use the Intersection over Union (IoU) between two events as the matching criteria. A detected event is considered as matching a ground truth event if their IoU score is larger than a threshold (typically 0.5). We also associate a confidence score with each detected event by simply averaging the single frame probabilities. This confidence allows us to characterize our detection output. We use average precision (AP) to evaluate the event detection. The AP describes the shape of the precision-recall curve, and is defined as the mean

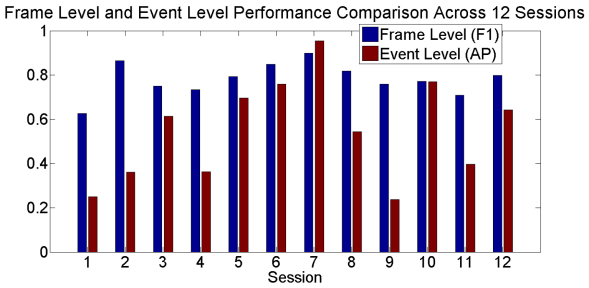


Fig. 8: The blue bars indicate frame level results (F_1 score), and the red bars indicate event level results (AP). Both frame level and event level results are based on the same PEEC+CRF output. There exists a large discrepancy between frame level and event level results in some sessions.

precision at a set of eleven equally spaced recall levels $[0, \frac{1}{m}, ..., 1]$:

$$AP = \frac{1}{m+1} \sum_{r \in 0, \frac{1}{m}, ..., 1} p(r), \tag{8}$$

where p is the precision, r is recall, and $p(r)$ is the measured precision at recall r. We set $m = 10$ similar to [7]. All results are reported using leave-one-subject out cross validation.

Our event detection results are obtained by PEEC+CRF output. We achieve an average AP of 0.5490 across the 12 sessions. We emphasis that event detection is significantly different from single frame detection. The former requires obtaining a continuous set of frames with its precise onset and offset, and the latter considers each frame independently. We plot the *same* PEEC+CRF output evaluated by frame level F_1 scores and event level AP in Fig. 8 for all sessions. There exists a large discrepancy between frame level and event level scores. A number of sessions (1, 2, 4, 9) have a fairly good F_1 score yet the corresponding AP is low.

In Fig. 9, we visualize PEEC+CRF output against the ground truth for one of the sessions with high F_1 and low AP. If we look at each frame separately, our method successfully picked most of the eye contact frames. If we look at the events, however, we failed to cover a large part of the events. We amplify a failure case in Fig. 9, where our method generates a small gap within a ground truth event and breaks it into two events. Due to low IoU scores, our evaluation criteria fail to match either of these events to the ground truth. We further sample a few frames within this event in Fig. 9. The gap is caused by partial occlusion of the face, which also results in failed face detection. Even a small amount of face detection failures can lead to a large discrepancy between frame level and event level results.

We also experimented with different IoU thresholds. As we vary the IoU threshold from 0.3 to 0.7 (Fig. 10), AP scores of both PEEC-based and PEEC+CRF-based event detection increase as the IoU threshold decreases. Moreover, Fig. 10 shows that our CRF model boosts the performance of event detection over all IoU thresholds, especially when the IoU threshold is high (IoU $= 0.7$).

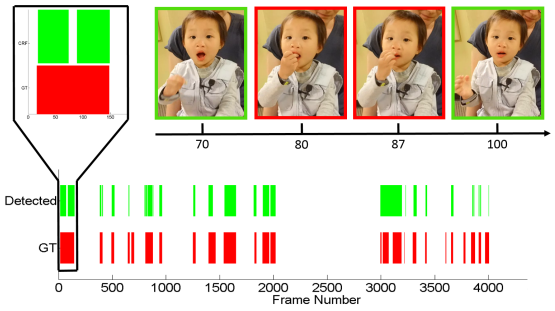


Fig. 9: Event visualization for session 2. Bottom row: comparison between detected events based on PEEC+CRF outputs and ground truth events. The red bars indicate ground truth events, and the green bars indicate detected events. Top row: Zoomed-in image of the gap between two detected events. Sample frames from the detected events (green bounding box) and from the gap between the two detections (red bounding box). The axis below the images indicates the corresponding frame numbers of these images.

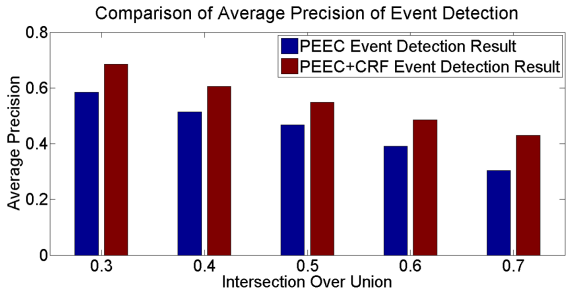


Fig. 10: AP of event detection with different IoU thresholds. The red bars are results based on PEEC+CRF, and the blue bars are results based solely on PEEC.

V. CONCLUSION AND FUTURE WORK

We consider the detection of bids for eye contact directed from a child to an adult who is wearing a POV camera during a naturalistic social interaction. We propose the first system that is capable of detecting the events of bids for eye contact in an egocentric video based on PEEC and CRF. We also present the first dataset of egocentric videos collected in the course of child-adult social interactions. The dataset includes 12 child subjects and will be made available to the research community. We provide a thorough evaluation of our method by comparing it to previous approaches. Our results outperform all other state-of-the-art methods by a large margin.

The major bottleneck of our method is the face detector in egocentric videos. Current face detection and alignment systems are not designed for child's face in an egocentric setting. We plan to develop a face detection and alignment pipeline tailored for egocentric videos, and extend our work to detect the child's looks toward objects, as well as the child's gaze shift between objects and eyes.

VI. ACKNOWLEDGMENTS

The authors would like to thank OMRON Corp. for providing OKAO Vision software. Portions of this work were supported in part by NSF Expedition Award number 1029679, the Intel Science and Technology Center in Pervasive Computing, and the Simons Foundation SFARI program.

REFERENCES

[1] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[3] K. Chawarska and F. Shic. Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39:1663–1672, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[5] 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. *Developmental Psychology*, 40(2):271–283, 2004.

[6] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085. IEEE, 2010.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 Results.

[8] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

[9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[11] T. Foulsham, E. Walker, and A. Kingstone. The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, 51(17):1920 – 1931, 2011.

[12] J. M. Franchak, K. S. Kretch, K. C. Soska, J. S. Babcock, and K. E. Adolph. Head-mounted eye-tracking of infants' natural interactions: a new method. In *ETRA*, pages 21–27, 2010.

[13] J. Guo and G. Feng. How eye gaze feedback changes parent-child joint attention in shared storybook reading? an eye-tracking intervention study. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2011.

[14] H.-C. Hsu, A. Fogel, and D. S. Messinger. Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. *Infant behavior and development*, 24(1):107–128, 2001.

[15] T. Hutman, M. K. Chela, K. Gillespie-Lynch, and M. Sigman. Selective visual attention at twelve months: Signs of autism in early social interactions. *Journal of autism and developmental disorders*, 42(4):487–498, 2012.

[16] W. Jones, K. Carr, and A. Klin. Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, 65(8):946–954, 2008.

[17] K. Kaye and A. Fogel. The temporal structure of face-to-face communication between mothers and infants. *Developmental Psychology*, 16(5):454, 1980.

[18] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[19] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 59(9):809–816, 2002.

[20] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. In *Archives of General Psychiatry*, 2002.

[21] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[22] M. Land and B. W. Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009.

[23] S. R. Leekam, B. Lopez, and C. Moore. Attention and joint attention in preschool children with autism. *Developmental Psychology*, 36(2):261–273, 2000.

[24] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013.

[25] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, 2011.

[26] D. S. Messinger, A. Fogel, and K. L. Dickson. What's in a smile? *Developmental Psychology*, 35(3):701, 1999.

[27] D. Model and M. Eizenman. A probabilistic approach for the estimation of angle kappa in infants. In *ETRA*, pages 53–58, 2012.

[28] P. Mundy and C. F. Acra. Joint attention, social engagement, and the development of social competence. In P. J. Marshall and N. A. Fox, editors, *The development of social engagement: Neurobiological perspectives*, pages 81–117. Oxford University Press, USA, New York, NY, US, 2006.

[29] B. Noris, M. Barker, J. Nadel, F. Hentsch, F. Ansermet, and A. Billard. Measuring gaze of children with autism spectrum disorders in naturalistic interactions. In *Engineering in Medicine and Biology Society, IEEE*, pages 5356 –5359, September 2011.

[30] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[31] J. Rehg, G. Abowd, A. Rozga, M. Clements, R. Mario, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Y. Zhefan. Decoding children's social behavior. In *CVPR*, 2013.

[32] A. Rozga, T. Hutman, G. S. Young, S. J. Rogers, S. Ozonoff, M. Dapretto, and M. Sigman. Behavioral profiles of affected and unaffected siblings of children with autism: Contribution of measures of mother–infant interaction and nonverbal communication. *Journal of Autism and Developmental Disorders*, 41(3):287–301, 2011.

[33] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, Portland, OR, June 2013.

[34] T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *ICPR*, 2014.

[35] A. Senju and M. H. Johnson. Atypical eye contact in autism: models, mechanisms and development. *Neuroscience and biobehavioral reviews*, 33(8):1204–1214, 2009.

[36] M. Sigman. The emanuel miller memorial lecture 1997: Change and continuity in the development of children with autism. *Journal of Child Psychology and Psychiatry*, 39(6):817–827, 1998.

[37] B. Smith, Q. Yin, S. Feiner, and S. Nayar. Gaze Locking: Passive Eye Contact Detection for HumanObject Interaction. In *ACM Symposium on UIST*, pages 271–280, Oct 2013.

[38] E. H. Spriggs, F. D. L. Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Egovision Workshop*, 2009.

[39] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR*, 2014.

[40] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *ECCV*, pages 656–667. Springer, 2008.

[41] A. M. Wetherby, N. Watt, L. Morgan, and S. Shumway. Social communication profiles of children with autism spectrum disorders late in the second year of life. In *Journal of Autism and Developmental Disorders*, 2007.

[42] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *CVPR*, 2013.

[43] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the UbiComp*, pages 699–704. ACM, 2012.

[44] W. Yi and D. Ballard. Recognizing behavior in hand-eye coordination patterns. In *International Journal of Humanoid Robots*, 2009.

[45] C. Yu and L. B. Smith. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, 8(11):e79659, 2013.

[46] L. Zwaigenbaum, S. E. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari. Behavioral manifestations of autism in the first year of life. In *International Journal of Developmental Neuroscience*, 2005.