Visualizing Gaze Direction to Support Video Coding of Social Attention for Children with Autism Spectrum Disorder

Keita Higuch[1] Soichiro Matsuda[2] Rie Kamikubo[1] Takuya Enomoto[3]
Yusuke Sugano[4] Junichi Yamamoto[3] Yoichi Sato[1]

[1]The University of Tokyo [2]Tsukuba University [3]Keio University [4]Osaka University

ABSTRACT

This paper presents a novel interface to support video coding of social attention in the assessment of children with autism spectrum disorder. Video-based evaluations of social attention during therapeutic activities allow observers to find target behaviors while handling the ambiguity of attention. Despite the recent advances in computer vision-based gaze estimation methods, fully automatic recognition of social attention under diverse environments is still challenging. The goal of this work is to investigate an approach that uses automatic video analysis in a supportive manner for guiding human judgment. The proposed interface displays visualization of gaze estimation results on videos and provides GUI support to allow users to facilitate agreement between observers by defining social attention labels on the video timeline. Through user studies and expert reviews, we show how the interface helps observers perform video coding of social attention and how human judgment compensates for technical limitations of the automatic gaze analysis.

Author Keywords

Video Coding Support, Children with ASD, Social Attention

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social communication and repetitive patterns of behaviors [3]. In 2016, Autism and Developmental Disabilities Monitoring (ADDM) Network reported that 1 in 68 children has been identified with ASD [13]. Children with ASD often face difficulties with daily functions.

Assessments of children with ASD are designed to measure and facilitate development of social communication skills such as imitation, pretend play, and joint attention. Therapeutic activities for developing these skills have been introduced into actual sites for ASD children's support. Recent studies revealed that the therapeutic activities can significantly affect their sociability in the long term [8, 33].

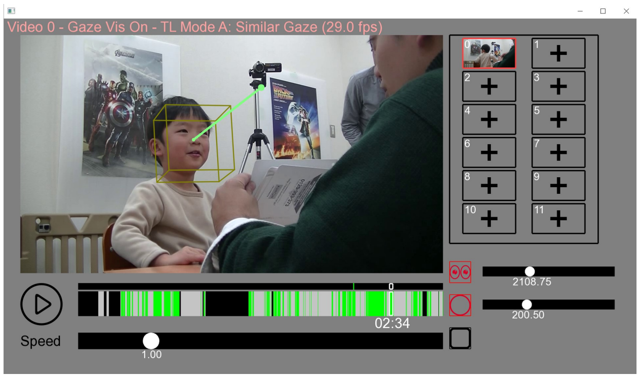

Figure 1. Proposed gaze visualization and video coding interface. The interface provides supportive features for video coding of social attention in assessments of children with autism spectrum disorder (ASD). Note that a child with typical development (TD) is shown in this picture.

Several structured therapeutic activities are supplemented with video recordings for evaluation. The evaluations are often performed as a post process of manual annotation of video recorded assessments [30, 6, 1, 2]. The video recordings are usually long and include many redundant scenes such as intervals of the session and breaks. Video coders (observers) have to locate scenes of assigned tasks in therapeutic sessions in lengthy videos and carefully extract attentive behaviors of children, such as eye contact and joint attention, for evaluating their social skills. Consequently, video coding of therapeutic activities requires a lot of effort from the observers. To make therapeutic activities more efficient, it is important to reduce the cost of video coding.

Partly because of such therapeutic demands, automatic analysis of attentive behaviors has been one of the central topics in the eye tracking community. However, despite recent advances in machine learning-based remote gaze estimation [48, 26, 49] and wearable camera-based eye contact detection [43, 12, 47], fully automatic detection of social attention under diverse environments is still challenging. Computer vision methods often suffer from various error factors, including diverse illumination conditions and occlusions, and we cannot fully rely on the automatic detection results. This issue becomes more critical in the case of therapeutic activities because incorrect evaluations must be avoided.

In this paper, we propose a supportive intelligent interface for video coding of therapeutic activities, as shown in Figure 1.

IUI'18, March 7–11, 2018, Tokyo, Japan

© 2018 ACM. ISBN 978-1-4503-4945-1/18/03. . . 15.00

DOI: https://doi.org/10.1145/3172944.3172960

We take an approach where we use state-of-the-art computer vision-based automatic video analysis to help guide human evaluation and judgment of social attention behaviors. The proposed interface visualizes the results of gaze estimation of target video images. Moreover, we provide users with features to define social attention labels in the videos in relation to the estimated gaze positions. Using sample-based or image region-based extraction methods, users can view candidates that indicate the parts of the videos with potential target social-attention behaviors such as eye contact.

The contribution of this study is threefold: 1) we interviewed six professional therapists for children with ASD as a preliminary study to identify difficulties in evaluating therapeutic activities with video recordings, 2) we designed a novel interface to support evaluation of lengthy videos to find target social attention of children with ASD, and 3) through user studies, we confirmed that the interface helps observers perform video coding of social attention, and we determined how it can be improved for evaluating therapeutic activities. We also discussed the impact of automatic detection errors in video coding tasks, and concluded that the proposed interface is still useful even with imperfect automatic video analysis.

RELATED WORK

Interactive technologies for Children with ASD

Interactive technologies for children with ASD have been developed in the context of Human-Computer Interaction (HCI) to care for and aid impairments such as social communication and repeated patterns [8]. These technologies for the assessment of social skills are aimed at developing reception to and production of social signals [10]. Successful interactive technologies have introduced tablet computers [29, 31, 20], special electronic devices [28, 38, 39, 9], and virtual avatars [19, 25, 46] for enabling therapeutic activities with information and communications technology (ICT). These computer-supported systems allow therapists to automatically code low-level behaviors of children with ASD [40, 38, 28].

With that said, therapeutic activities in face-to-face settings with expert therapists are still important to assess and develop social skills of children with ASD. To accurately evaluate, therapists frequently perform video-based coding with recorded videos during therapeutic activity [30, 6, 1, 2]. However, video coding presents several difficulties in marking social attention in long video recordings. To address these difficulties, this work aims to support video coding of therapeutic activity for children with ASD by visualizing computer-estimated gaze direction.

Intelligent User Interfaces for Video Annotation and Browsing

Previous work developed interfaces to support video annotation (coding) of behaviors for manual annotation by observers [42, 17]. To reduce the effort of video coding, previous work introduced automatic behavior detection features such as sign language [15], transcription of speech [7], and facial expression [16]. Unlike these works, we introduce gaze analysis of videos of children with ASD to offer video coding support to experts. To this end, we performed interview studies with experts and designed supportive features to find target social attention from long video recordings of therapeutic activities. The novelties of the features are twofold: 1) introducing automatic gaze analysis to help users to detect attention behaviors, and 2) allowing users to define social attention labels for highlighting candidates of target social attention in video coding tasks.

Recent works also aimed at developing new tools for efficiently browsing diverse videos such as education [24], surveillance [36, 35], sports [27], and first-person videos [18]. These works used computer vision techniques and developed novel GUI supports for video browsing such as direct manipulation [32, 21, 22], content-aware fast-forwarding [18, 11, 34], and a colored video timeline [18, 14]. The colored timeline highlights part of a video in the timeline based on automatic detection results to indicate cues of important scenes to observers. Despite the benefits of the previous interface, it only used static data from preprocessing for the highlights. In this work, we developed a new interactive colored timeline feature of which highlighted parts change based on real-time user input.

Detecting and Analyzing Social Attention

With the progress in machine learning methodologies, automatic attention analysis has significantly improved. State-of-the-art face and facial landmark detection methods [4, 5, 41] are very robust even in challenging conditions, and calibration-free, appearance-based gaze estimation methods have advanced with large-scale datasets and deep learning techniques [48, 26, 49]. Although such learning-based methods have an advantage in that they require only a consumer RGB camera, they still suffer from various error factors, and the accuracy (average error: 5 degrees) is far less than commercial eye trackers. For this reason, we use a computer vision method with the GUI in video coding therapeutic activities.

Recent studies also aimed at automatically detecting events of social attention, such as joint attention [23, 45] and eye contact [43, 37, 44, 12], with wearable cameras or eye trackers. However, there are several fundamental difficulties in detecting attentive behaviors with wearable devices. First, it is difficult for children to wear cameras or eye trackers, and the approach is not always applicable in practical situations. Second, as discussed in [47], the definition of eye contact depends on the target object, and it is almost impossible to *pre-train* robust eye contact detectors without knowing the target object. Finally, social attention also highly depends on the context of interaction. Eye contact in social interaction has multiple meanings [30], and observers need to understand the meaning of social attention. These difficulties also motivated us to use automatic video analysis in a supportive manner for guiding human judgment.

DESIGN CONSIDERATION

Preliminary Study

To design a support interface for video coding, we first performed interviews of six experts (two female, four male) of children with ASD. The experts have over five years of experience with clinical assessments and academic studies. The semi-structured interview sessions took place in a private room at their facility, where we asked the following three questions: 1) In what situations do you use video recording for evaluations of assessments? 2) Are there any difficulties with the video coding–if so, what? 3) Do you want to use new technologies for video scoring–if so, what kind of support is crucial?

In addition to the interviews, we also observed how the experts performed the video coding. The experts showed us videos of assessments of the developing social skills of children with ASD according to the ethical code of our institute. They also showed their video coding tools [1] and described how they perform evaluations of therapeutic activities from the recorded videos.

Use of Video Recordings

According to the interviews, the experts frequently use video recordings for evaluations of assessments of children with ASD. Typically, target behaviors such as eye contact, joint attention, and imitations are clearly defined for each assessment session. Target behaviors can appear multiple times in each recording session, and the experts need to repeatedly review the recordings to code correctly.

There are two levels of video coding children's behaviors. The first level is identifying large actions and reactions of a child, such as pointing and imitations. These target behaviors are often clear in the recorded videos, and observers can easily confirm whether a child did the target actions/reactions. The second level is identifying more complex attentive behaviors, including eye contact (face looking) and joint attention with the therapist. Such attentive behaviors are important to measure the social skills of children.

The experts are generally welcome to use new tools to reduce their efforts and improve their productivity. However, it is not always possible to use hardware support (such as eye trackers) in face-to-face settings. This is one of the main reasons why the experts rely on video coding. Since software support tools for video coding are still quite limited, they sometimes recruit undergraduate students as assistants for video coding.

Difficulties of Video Coding

There is a fundamental ambiguity in finding attention in 2D video observations. Observers need to carefully examine the facing and gazing directions of children to predict their attention. This requires a huge effort for observers to identify target attentive behaviors particularly in the case of complex social attention.

In addition, video recordings are usually long (10 to over 60 minutes) because they continuously capture videos of entire therapeutic activities that include both on-task and off-task (*e.g.*, break) scenes. Therefore, observers spend a long time locating on-task scenes in long videos, and it is sometimes difficult to find all of the target behavior instances in the videos.

Our Approach

The main goal of this work is to address the above difficulties of video coding social attention in face-to-face settings (Figure 2). We take the approach of using recent advancements in computer vision technologies that allow us to detect attention behaviors (*e.g.*, gaze directions) from videos. However, robust, fully automatic recognition of social attention is still difficult due to limited accuracy and the ambiguity of attention (*i.e.*, attention dependent on contexts of social interaction) [49, 30]. This limitation motivated us to design a new approach using automatic video analysis to support guiding human judgment for facilitating agreement between observers.

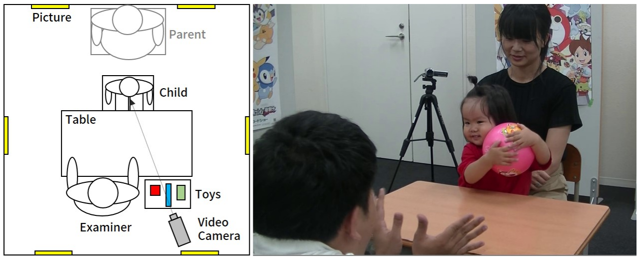

Figure 2. Example of target setting: assessments of children with ASD in a face-to-face setting (redrawn based on [30])

Considering the underlying difficulty and ambiguity of social attention judgment, we propose that gaze direction for video coding helps observers' judgments of social attention. We expect the visualization facilitates agreement between observers. In addition, to address the difficulty in finding multiple targets from a lengthy video, our interface provides supportive features that highlight candidates of targets based on user-defined social attention labels. Following prior work that reported the benefits of colored timelines (*i.e.*, [18, 14]), we also propose highlighting detection results in the video timeline of the interface. These features narrow down candidate frames of social attention in long video recordings and are expected to reduce observers' efforts.

PROPOSED INTERFACE

Based on the preliminary study, we propose a novel video coding interface that provides supportive features using computer vision techniques. Figure 3 shows the overview of the proposed framework. As a preprocessing step, the face and gaze direction of the target child is automatically detected in each frame of the input video. The proposed interface uses these detected data to visualize gaze directions.

Gaze Analysis

We use the OpenFace toolkit [5] to detect and track the target child's face. The toolkit provides 2D facial bounding boxes, facial landmark locations, and 3D head poses. Since the input video often contains many non-target faces, such as those of parents, our interface allows observers to define a target region of interest (ROI) in the input video frames. The ROI indicates the working area of children during therapeutic activities, and our interface ignores faces detected outside the ROI.

After face detection, the state-of-the-art appearance-based gaze estimation method [49] is then applied to estimate gaze directions of detected faces. 3D head poses are used to crop normalized face images and the face images are fed into a pre-trained full-face convolutional neural network for gaze

[1] Microsoft Media Player with Microsoft Excel, BECO2 and Noldus Observer

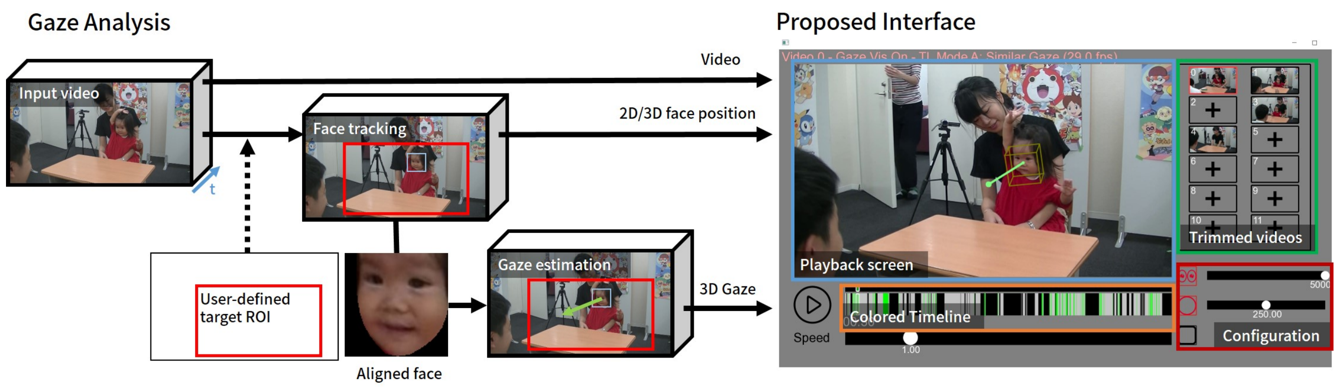

Figure 3. Overview of gaze analysis and the proposed interface: Gaze analysis takes face tracking and gaze estimation stages (Left). The proposed interface provides basic functions of video coding and supportive features for finding target social attention (Right).

estimation. Assuming that the camera's intrinsic parameters were obtained through camera calibration, this process provides the 3D gaze vectors of the target child defined in the camera coordinate system.

Interface Design

The right part of Figure 3 shows the proposed interface, which contains basic functions of video coding tools such as the playback screen, the video timeline, and the configuration panels. Parts of the video timeline can be highlighted based on detection results. The interface also has an annotation slider on the top of the timeline. Observers are allowed to mark on the slider by pressing a key to manually annotate when they find target social attention.

Visualizing Gaze Analysis Results

The interface also shows results of detected information on video images. We visualize both 3D face boxes and 3D gaze vectors projected onto video images. Proper lengths of gaze vectors differ by face positions and the context of the videos. The interface thus provides a slider to adjust the lengths of gaze vectors. The interface draws gaze vectors from the center of the face to the gaze position of the selected length.

The interface also shows face detection results on the video timeline. Previous studies revealed that highlighting analysis results in video timelines is beneficial for finding specific information from videos [14, 18]. We introduce this benefit into the video coding interface. As shown in Colored Timeline of Figure 3 (right), the interface visualizes face detection results in the video timeline. Gray regions indicate that a child's face appeared in a defined detection region. The black region shows that no face is detected in the region. This visualization was designed to help observers guess when a child was working on given tasks during video recordings.

Supportive GUI Features for Finding Target Social Attention

The interface also provides supportive features to interactively highlight candidates of target behaviors based on user-defined social attention labels. To define target labels, we propose two types of selection methods. The first is an sample-based method that allows observers to select samples of gaze direction as a target social attention label. Candidates of the target are then extracted and highlighted based on a distance-based search. The second is a region-based method that allows observers to draw a region on video frames. Candidates of the target are then highlighted if visualized gaze points appear inside of the target region. These methods work as GUI supports of the proposed interface. To address errors of gaze estimation results, the approach enables interactive adjustments of a detection threshold or region.

Sample-based highlight allows users to select 3D gaze positions as samples of their observational targets of social attention. Such user-defined labels are used to highlight frames detected based on the 3D distance between the center of sample gaze points and other gaze points in each frame. The gaze points are the tips of the gaze vectors whose lengths observers freely set from the face positions. The interface also allows the users to define a threshold of the distance for highlighting frames in the video timeline as the candidates.

Figure 4 (A) shows a process of using the sample-based highlight. In this mode, observers need to find the first sample of target social attention and mark it on the interface (A1). The interface then highlights detected candidates as green on the timeline (A2). The gaze vector is also highlighted as green if the seek bar of the timeline is set on detected frames. Observers can change the detection threshold of candidates (A3) and add further target labels (A4).

Region-based highlight allows users to set a rectangle on the video images as a target region of social attention. In the frames where the projected points of gaze vector tips on the playback screen are detected within the user-selected region, the interface highlights those frames on the timeline as the candidates. Unlike the sample-based highlights, the users can select the target regions multiple times and the highlighted candidates are updated according to the newly selected region.

Figure 4 (B) shows a process on using the region-selected highlight. Observers need to set a target region on the playback screen to highlight candidate frames by dragging the mouse (B1). In the selection process, green points on

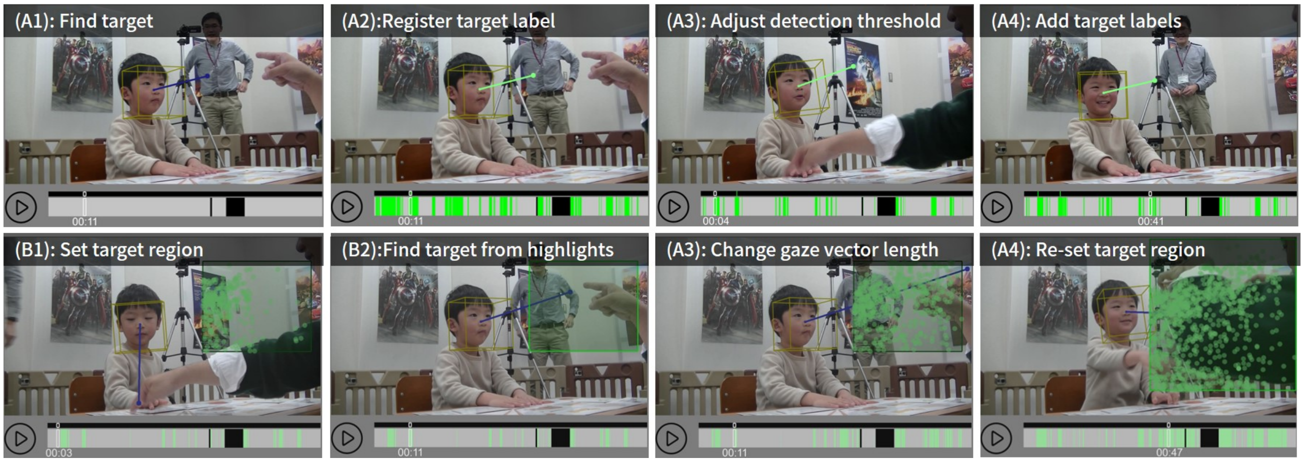
(A1): Find target
00:11
(A2):Register target label
00:11
(A3): Adjust detection threshold
00:04
(A4): Add target labels
00:41
(B1): Set target region
00:03
(B2):Find target from highlights
00:11
(A3): Change gaze vector length
00:11
(A4): Re-set target region
00:47

Figure 4. GUI supports for highlighting candidates of target social attention: (A1-4) Sample-based highlight, and (B1-4) Region-based highlight.

the screen indicate the detected gaze tips of each frame. The detected frames are also visualized on the video timeline. The rectangle of the selected region is highlighted if the seek bar of the timeline is set on detected frames (B2). Observers are allowed to change the length of the gaze vector (B3) and target regions (B4) based on their preference.

Video Trimming Function

The interface supports interactive video trimming for long videos. Video recordings of assessments are usually long and contain on-task scenes in limited parts. Observers spend a lot of effort finding and repeatedly checking the scenes in video coding. To reduce this effort, this feature aims to support observers in easily trimming important scenes from long videos. To trim the video, observers select a region in the timeline by dragging the right button of a mouse. Observers then click rectangles on the right side of the interface to make a new trimmed video. We expect the trimming function to reduce the effort of repeatedly finding and playing on-task scenes.

DATA COLLECTION FOR THREE EVALUATIONS

We created a novel video dataset of therapeutic activities for both children with typical development (TD) and ASD. We designed three evaluation studies for the visualization of gazes and the proposed interface. The first evaluation is confirming the effectiveness of visualizing gaze estimation results by a simple task. This task is designed to see whether the visualization helps observers with video coding of social attention, more specifically face looking of children, even without GUI supports. The second evaluation is revealing how the proposed features of the interface are used in a simple video coding task such as usability. We employed the baseline coding interface with two variations of the proposed features. We recruited both non-expert and expert therapists for the evaluation to understand the usability of the proposed interface with enough numbers of participants. Note that the usability for non-experts is also important because, in our interviews, experts mentioned that undergraduates are sometime recruited for video coding of assessments. We finally performed expert reviews in which experts use the proposed interface for video recordings of realistic therapeutic activities. We aimed to see how the proposed features can be used for our target of video coding.

Dataset

We recruited three expert therapists to perform professional therapeutic activities. We also recruited two children with TD (4-year-old boy and 2-year-old girl) and two children with ASD as targets of assessments. According to the ethical code of our institute, we used video recordings of children with ASD only for the evaluation of expert reviews.

We recorded scenes of therapeutic tasks assigned to children (Figure 5). More specifically, the experts performed imitation and naming tasks with the children. The imitation task is designed to see whether a target child is able to imitate several actions of the therapists. The naming task is used to see development of social interaction. In the task, a therapist shows or points to a picture card for the target child to look at and respond to by naming what is in the picture (e.g., animals). We captured video recordings for both children with TD and ASD during the social interaction tasks (the first row of Figure 5).

We also recorded a structured assessment program to evaluate the effectiveness of the proposed interface with actual therapeutic activities. We selected the early social communication scales (ESCS) [30], a video-recorded, structured observation measure to detect social disabilities in children. More specifically, ESCS was designed to provide measures of individual differences in nonverbal communication skills that typically emerge in children between 8 and 30 months of age. ESCS' settings (Figure 2) and nine tasks that are designed to detect specific social abilities such as joint attention, social interaction, and behavior requests are clearly defined. We simplified the use of ESCS to reduce the effort of children by choosing four typical tasks. We captured video recordings during the ESCS sessions with two children with TD (second and third rows of Figure 5).

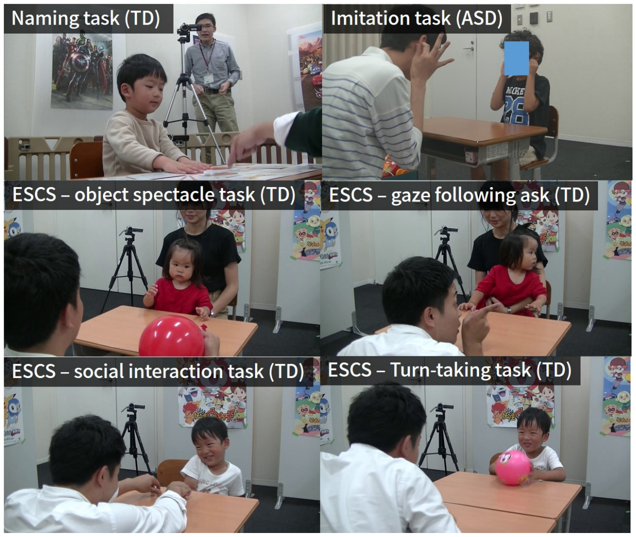
Naming task (TD)
Imitation task (ASD)
ESCS – object spectacle task (TD)
ESCS – gaze following ask (TD)
ESCS – social interaction task (TD)
ESCS – Turn-taking task (TD)

Figure 5. Our video dataset. The dataset contains general assessments (the first row) and a structured measurement program (the second and third rows) for both children with typical development (TD) and autism spectrum disorder (ASD). To follow the ethical code of our institute, the face of the child with ASD is hidden in the picture

EVALUATION 1: EFFECT OF VISUALIZING ESTIMATED GAZE

We first aimed to confirm our basic assumption that visualizing gaze direction helps to detect social attention activities. To this end, we used a single image to perform a simple study in which participants judged whether a child made face-looking behavior or not. We compared visualized and non-visualized conditions to see how visualized gaze direction changed observers' judgments. We recruited 12 participants who were non-experts of children with ASD but had enough experience with computers in their studies and works.

Task

Participants were asked to code whether a child on a given image looked at a therapist's face (face looking behavior). We requested that the participants quickly and accurately judge the face looking of a target child. The participants simply selected an answer from two keys (yes/no) for 100 images.

From our dataset of children with TD, we captured 50 images of scenes involving face looking behaviors performed by a target child and 50 images for non-face looking behaviors. The set of images had a variety of children, therapists, situations, and fixation targets in non-face looking behaviors. We shuffled the order of 100 images and randomly labeled half of the images with 0 and the other half with 1. Half of the participants saw images of 0 labels with visualized gaze and images of 1 labels without gaze visualization. The other half had reversed conditions.

Evaluation and Results

We compared answers for visualized images with answers for non-visualized images in terms of agreement ratio and response times. For the agreement ratio, we compared participants' selections with our selection for visualized and non-visualized conditions. We also calculated average response times for the conditions. As part of the participants' objective feedback, we also asked participants to identify in which conditions (*i.e.*, visualized and non-visualized conditions) was it easy to judge face looking after their tasks.

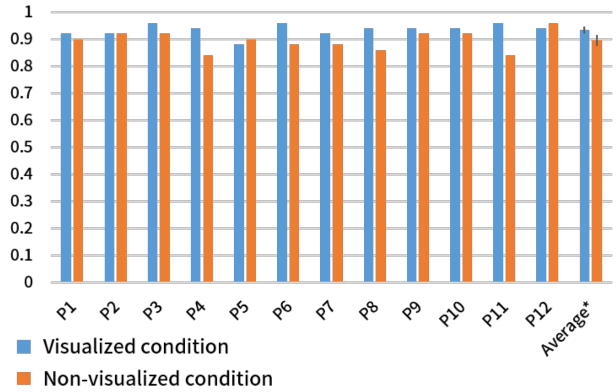

Figure 6. Result of agreement ratio. The Mann-Whitney U test revealed significance for the average agreement ratios ($p = 0.004$). Error bars indicate 95 % confidence interval.

Figure 6 shows the results of agreement ratios for each participant. Average agreement ratios of visualized and non-visualized conditions are 0.93 (SD: 0.03) and 0.89 (SD: 0.04), respectively. The Mann-Whitney U test revealed the significance of the agreements ratio ($p = 0.004$). Average response times for visualized and non-visualized conditions were 2.00 seconds (SD: 0.34) and 2.00 seconds (SD: 0.38), respectively, so no significance was observed by the Mann-Whitney U test. In terms of objective feedback, all participants agreed that it was easy to judge face looking in the visualized condition. Despite there being no difference in the response times, we confirmed the assumption that the visualized gaze direction of a child helps observers facilitate agreements for coding social attention.

EVALUATION 2: INTERFACE USABILITY

We performed a study to evaluate the proposed features in addition to the overall functions of our system. We aimed to see how the experience of searching for target social-attention behaviors from therapeutic activity videos was affected by our system. We recruited two types of participant groups, non-experts (n = 12) and experts (n = 5), in video coding and observation of social attention. They were asked to complete three sets of tasks under different conditions in using the proposed interface features.

Conditions

The participants completed the tasks under the counterbalanced conditions, which were as follows:

Baseline (None)

Under this condition, they performed the assigned tasks without the visualization of gaze estimation, highlight features of face-detection results, or user-defined attention labels. The participants were given only the basic functions of video coding tools, including the playback screen, the video timeline, and the speed configuration. The interface had an annotation slider with which they could position a mark by pressing a key.

Proposed User-Defined Highlight: Sample-based (Sample)

While the basic functions described above were provided, under this condition the participants were expected to utilize the sample-based highlight features. They were instructed to sample out target social attention by marking it on the annotation slider to signify highlights of closely detected candidates on the timeline. The condition came with a slider handle to control the threshold of the distance of the 3D gaze positions at frames marked in relation to other frames.

While activating one of the user-defined highlight features, the visualization of gaze estimation and the face-detection highlights were always available. The participants could freely manipulate the length of the stick for gaze direction using the slider handle.

Proposed User-Defined Highlight: Region-based (Region)

Under this condition, the participants were expected to utilize the region-based highlight features. They were instructed to select a rectangular region on the video content, and the frames that included gaze positions detected within the region were highlighted.

Task and Procedure

The participants were given 3-minute-long video clips from the therapeutic activity recordings to search for scenes in which the child looked at the face of the therapist in the assessments. We assigned the participant to code the target behaviors within three-second intervals. The clips we used were from the data collected in the recordings of the 4-year-old boy with TD. The participants were asked to annotate a mark on the slider located on top of the timeline every time they found the target scenes. They performed these manual annotation tasks of finding target social attention behaviors under different usage of the interface features described above. Prior to each usage condition, a practice session was provided to allow the participants to get familiar with the interface.

After the experimental sessions, the participants were also given Likert scale survey questions to rate the ease of tasks under each condition and the effectiveness of features for *Sample* and *Region* conditions. The study then proceeded to interviews to understand the reasons for their ratings in the context of use and to receive their subjective feedback towards the proposed features of our system.

Quantitative Evaluation

We calculated the total averages of agreement ratios between all of the participants in each video clip. The results of *Sample*, *Region*, and *None* are 88%, 86%, and 86%, respectively. We also compared average completion times. The results of *Sample*, *Region*, and *None* are 346.98 (SD:123.94), 363.31 (SD:97.68), and 293.78 (SD:92.87). The Friedman test revealed no significant difference within the conditions. Note that these results are not surprising because the video recordings used in this evaluation are short in length (only 180 seconds), and the participants were requested to manipulate several GUI components relevant to the proposed features.

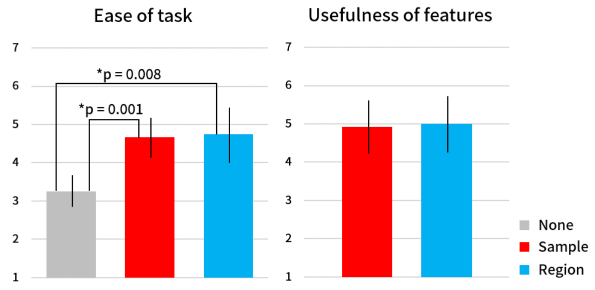

Figure 7. Results of objective feedback for the ease of tasks and usefulness of features from 12 non-experts. The Friedman and Mann-Whitney U tests revealed significance in the ease of tasks for *None-Sample* ($p = 0.001$) and *None-Region* ($p = 0.008$). No significance was observed in the usefulness of features from the non-experts. Error bars indicate 95 % confidence interval.

In the objective feedback from 12 non-experts, given that the averages regarding the ease of tasks in *Sample*, *Region*, and *None* were 4.67 (SD:1.07), 4.75 (SD:1.54), and 3.25 (SD:0.87), respectively, the Friedman test revealed significance within conditions (left of Figure 7). The Mann-Whitney U test revealed significance in *None-Sample* ($p = 0.001$) and *None-Region* ($p = 0.008$). Among five experts, the average answers for *Sample*, *Region*, and *None* were 5.2 (SD:1.10), 5.4 (SD:1.52), and 5 (SD:2.12), respectively. Given that the average answers from non-experts regarding the usefulness of features for *Sample* and *Region* were 4.92 (SD:1.51) and 5 (1.60), no significance was revealed by the Mann-Whitney U test (right of Figure 7). The answers from experts were 5.4 (SD:0.55) and 5.2 (SD:1.48). Overall, we observed positive objective feedback for two of the proposed features.

Subjective Feedback

We generally received positive feedback towards our system, as it was helpful in quickly estimating which portion of the timeline had information relevant to the completion of the task. They viewed the timeline highlights as an effective means of understanding the overall structure of the video content. The participants from both groups reported the highlighted candidates as: *"clues to determine the search portion in the videos."*[2]

Highlights on the timeline were found to be useful in estimating the sections participants needed to pay attention to. Many non-expert participants mentioned the fatigue level involved in not using the highlight features. One stated: *"The task itself becomes tiring. I had to constantly focus on the video."* In having no highlights, the experts mentioned how they were able to perform observational tasks at different attention levels. One expert participant who often performs video coding for counseling purposes reported: *"I became careful in observing*

[2] In this paper, italic fonts in double quotations denote translated speech from other languages.

the video when approaching the highlighted frames." We also observed that they often configured the speed slider to slow down the video or to move frame by frame to check around the estimated target frames. They added: *"I relaxed my attention while watching the content during dark-highlighted frames. That was not stressful."* Both groups of participants strongly emphasized the importance of the face-detection information in roughly estimating unnecessary sections of the content.

The majority of the participants preferred the use of sample-based highlights because they controlled system feedback by having visualized information (their markings on the video). Even though many reported cases of low accuracy of the system feedback, their acceptance of use was motivated by their flexibility in changing the thresholds to see the relationships between user-marked frames and others. Three non-expert participants articulated: *"I relied on my eyes to judge in the end"* in order to determine whether what they observed was valid. One expert explained: *"I can choose the relevancy of the candidates given on my own."*. On the other hand, few participants expressed a preference for the region-based highlights due to the lack of ease-of-use elements. They had to constantly change the parameters of the visualized gaze direction stick and re-select the region throughout the video content.

Among the participants, we observed a tendency to run through the whole video at least twice, and the highlights were then utilized to check the quality of their work annotating target behaviors. Without the highlights, non-expert participants observed: *"It took a lot of effort to look over the work without any guidance."* In comparing their work with the system-offered candidates, an expert participant said: *"Seeing how my annotated frames matched the highlights at the end increased my confidence level in my work."* In addition to minimizing coding errors, the participants also checked if they missed any relevant frames by going through the candidates. Without the highlight features, no clues were given on the timeline to effectively run through the frames all over again.

EVALUATION 3: EXPERT REVIEWS

This section explores the application of our system in the actual practice of video-based evaluations of social attention. We assigned five experts to review the interface under a given sets of tasks for video coding. In order to set practical conditions for using the interface for video coding, the study used a 3-minute-long clip from the therapeutic activity recording of the ASD child in the first session and proceeded with the whole 12-minute recording of the activities done by the two-year-old girl with TD. Rather than emphasizing multiple dataset scenarios with diverse children groups, we followed the ESCS coding tasks, which were generally performed on children with TD and ASD, and qualitative methods to gain thorough implementation aspects of the interface.

Task and Procedure

The task of the first session was to freely utilize the features (including sample- or region-based highlights, along with gaze-estimation and face-detection results visualized) to search for frames with target social attention as given in the interface usability study. It was followed by a session using our interface to perform simplified ESCS coding tasks for our expert-review purposes. In this scenario, the experts carefully observed and analyzed the long recordings for target attributes that fit behavioral categories structured by ESCS forms. We conducted unstructured interviews after each session to gain natural reviews of the features of our proposed interface and possible extension of the function provided in support of the complex and high-effort tasks of video coding.

The experts were instructed to trim video segments based on the ESCS coding tasks using our video-trimming function and to fill out our simplified version of ESCS forms. The example coding measures included a child's eye contact behaviors while playing with a given toy or when alternating their gaze between a moving toy and the examiner (in the category of initiating joint attention), requesting a toy (in the category of initiating behavioral requests), or interacting with the examiner in some way, such as singing (in the category of responding to social interaction).

Subjective Feedback

When analyzing the social-attention behaviors of children with ASD, all of the experts articulated how they made far less eye contact (face looking) and how this nature influenced the implementation reliability. Two experts mentioned that the sample-based timeline visualization for eye contact was particularly reliable in this task. One of them said: *I was amazed to see how the system was able to detect those few frames of limited social-attention behaviors."*

Since the clips from the ASD-child recordings involved content in which the child was mostly on the seat, one expert extended the use of the region-based highlights to select an area where there would not be any gaze direction or positions. This was found to be an effective way to visualize which portion of the video had unnecessary content.

The experts mentioned the use of timeline visualization features when trimming video clips. It was interesting to see how the system was implemented in the actual context of video coding. Face-detection was especially useful for long videos as a way to know which sections had unnecessary content such as a child being out of the frame. Moreover, the sample-based highlights gave hints as to which video segments most likely had experimental tasks. Since ESCS experiments usually come with strict procedures, the clipping feature was more useful in a video involving a child with ASD when it was difficult to estimate how the procedures would go. Knowing the overall structure of the video beforehand explicitly influenced the effectiveness of the highlights on the timeline.

FINDINGS

Benefits of Visualizing Gaze Estimation

Visualization of gaze estimation results on the video images helps video coding to facilitate agreements for assessing social attention. As shown in results in evaluation of visualizing gaze directions, this can be seen as minimizing the variability of judgment due to personal interpretation. The visualization can thus encourage the confirmation of target behaviors for the observers, whether they observed eye contact in the video frame.

Visualized gaze can be also used to define the labels necessary for highlighting candidates for target social-attention behaviors. For sample-based highlights, the observers search and select the representative line of sight from the visualized gaze directions. To make use of the region-based highlights, the observers perceive how the gaze directions are related to the defined areas of attention.

Benefits of Highlighting Detection Results over the Timeline

Using our detection methods to visualize candidates on the timeline supports observers in narrowing down parts of the videos to find target social attention. Face-detection highlights provide the initial support in understanding the general structure of the video content. The redundant parts of the video, such as when the child is off the seat or outside the camera's field of view, can be easily determined. In addition, the highlights related to gaze estimation come in handy when looking for parts that contain potentially relevant content. Using the sample-based features or region-selected features allows the observers to further narrow down the search with the ability to select the levels of candidate information they prefer to see on the timeline slider.

The benefit of reducing the search levels is that the cognitive load of complex observation is relieved because the observers know where to focus while the video is running. Without the highlight features, the majority of the participants mentioned how they had to remain attentive throughout the entire video. The video-coding experts reported ease in grabbing parts of the video that involved content that was outside the scope of the observational task. Expert feedback stated that the timeline highlights significantly minimized the effort in analyzing the video for potential frames with target social attention behaviors.

Considerations for Detection Error

It is important to discuss the impact of automatic detection errors. The accuracy of gaze estimation technology is still limited, and the analysis results of the datasets used in the evaluations revealed several types of errors. The face of the two-year-old child could not always be detected; the face detection results in the images of that child were not as stable as those from the four-year-old child's images (Left of Figure 8). Gaze estimation bias was also observed due to the occlusion of children's faces and the lack of relevant face images in the learning datasets (Right of Figure 8). In addition, the experts mentioned certain behaviors that could be uniquely found in ASD children, and the example case was of tilting the head downward while rolling eyes upward. Even though we did not observe those behaviors in the dataset we used, the experts reported the expected negative aspects considering various behavioral factors impacting the gaze estimation results.

Regarding the video coding performance from the ASD child experiments, the experts reported that they felt more confident in their own observational skills rather than relying on the system feedback. This was due to the fact that ASD children seldom make easy-to-define social attention behaviors. For instance, the expert observers were able to tell that the child looked at objects or somewhere else in the environment even though their eyes seemed to be directed towards the face of the examiner.

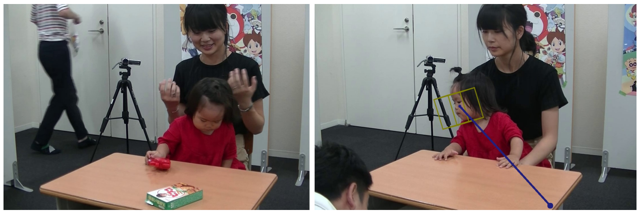

Figure 8. Failure cases of face detection (Left) and gaze estimation (Right)

Despite several detection challenges, the proposed interface along with the special features was still useful in finding target social attention. Since the observers relied on their own observational skills in making final judgments of target behaviors, their feedback on the use of the interface was positive. Simply offering the possibilities of relevant parts gave them general directions of where to focus in the long videos. The functions where the observers could manipulate the levels of possibility feedback met their search needs as well. They could choose to see limited numbers of options with higher-level relevancy and then expand more options to browse for potentially relevant frames. In support of the complex observational tasks, rather than automating them completely, our interface encouraged the observers to effectively utilize their human skills by being able to see which parts needed to be browsed quickly or thoroughly.

LIMITATION

The main limitation of this work is that we evaluated the proposed interface with limited variations of videos in the user studies. Video recording of therapeutic activities are usually long, but we only used videos 180 seconds long in the interface usability study (Evaluation 2). Even though primary findings of the proposed features were revealed in the study, we plan to perform additional user studies with realistic video lengths. The evaluations also only used the videos of therapeutic activities recorded in a face-to-face setting. Therefore, we plan to collect a new dataset of other settings (e.g., floor setting), and evaluate the proposed interface to reveal further requirements for improving the supportive features.

CONCLUSION AND FUTURE WORK

This paper presents a novel interface to support video coding of social attention for the assessments of children with ASD. Based on interview study with professional therapists, we designed our interface to use computer vision-based automatic video analysis in a supportive manner for guiding human judgment.

We performed three evaluations to find the effectiveness of the proposed interface. We found benefits of gaze visualization and supportive features for finding target social attention. Through evaluations, we also found that, through our proposed interface, human judgment often compensates for the technical failure of the automatic gaze analysis. Although computer vision-based gaze estimation still has some technical limitations, the proposed interface could make assistance for complex video coding tasks.

One of our most important future works is extending the proposed interface for use with multiple video sources. In the interview sessions, an expert mentioned that they sometimes use multiple videos for coding assessment recordings. While multiple video sources can lead to more accurate and detailed analysis of social attention, it will open new challenges for efficient visualization. We also plan to explore supportive features with more advanced automatic attention analysis approaches and perform more deployment studies during actual therapeutic activities.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR14E1 and JPMJCR14E2, Japan. We thank all collaborators in our user studies and dataset for their support.

REFERENCES

1. Catherine Aldred, Jonathan Green, and Catherine Adams. 2004. A new social communication intervention for children with autism: pilot randomised controlled treatment study suggesting effectiveness. *Journal of Child Psychology and Psychiatry* 45, 8 (2004), 1420–1430. DOI: http://dx.doi.org/10.1111/j.1469-7610.2004.00338.x
2. Nadig AS, Ozonoff S, Young GS, Rozga A, Sigman M, and Rogers SJ. 2007. A prospective study of response to name in infants at risk for autism. *Archives of Pediatrics Adolescent Medicine* 161, 4 (2007), 378–383. DOI: http://dx.doi.org/10.1001/archpedi.161.4.378
3. American Psychiatric Association and others. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
4. T. Baltrusaitis, P. Robinson, and L. P. Morency. 2013. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In *2013 IEEE International Conference on Computer Vision Workshops*. 354–361. DOI:http://dx.doi.org/10.1109/ICCVW.2013.54
5. T. Baltrusaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10. DOI: http://dx.doi.org/10.1109/WACV.2016.7477553
6. Grace T. Baranek. 1999. Autism During Infancy: A Retrospective Video Analysis of Sensory-Motor and Social Behaviors at 9–12 Months of Age. *Journal of Autism and Developmental Disorders* 29, 3 (01 Jun 1999), 213–224. DOI: http://dx.doi.org/10.1023/A:1023080005650
7. Ronald Böck, Ingo Siegert, Matthias Haase, Julia Lange, and Andreas Wendemuth. 2011. ikannotate–a tool for labelling, transcription, and annotation of emotionally coloured speech. *Affective Computing and Intelligent Interaction* (2011), 25–34.
8. Sofiane Boucenna, Antonio Narzisi, Elodie Tilmont, Filippo Muratori, Giovanni Pioggia, David Cohen, and Mohamed Chetouani. 2014. Interactive Technologies for Autistic Children: A Review. *Cognitive Computation* 6, 4 (01 Dec 2014), 722–740. DOI: http://dx.doi.org/10.1007/s12559-014-9276-x
9. LouAnne E. Boyd, Xinlong Jiang, and Gillian R. Hayes. 2017. ProCom: Designing and Evaluating a Mobile and Wearable System to Support Proximity Awareness for People with Autism. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 2865–2877. DOI: http://dx.doi.org/10.1145/3025453.3026014
10. Laurence Chaby, Mohamed Chetouani, Monique Plaza, and David Cohen. 2012. Exploring multimodal social-emotional behaviors in autism spectrum disorders: an interface between social signal processing and psychopathology. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 950–954.
11. Kai-Yin Cheng, Sheng-Jie Luo, Bing-Yu Chen, and Hao-Hua Chu. 2009. SmartPlayer: User-centric Video Fast-forwarding. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 789–798. DOI: http://dx.doi.org/10.1145/1518701.1518823
12. Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. 2017. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 43 (Sept. 2017), 20 pages. DOI:http://dx.doi.org/10.1145/3131902
13. DL Christensen, J Baio, and K Van Naarden Braun. 2016. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years – Autism and Developmental Disabilities Monitoring Network. *MMWR Surveill Summ* 65, SS-3 (2016), 01–23. http://dx.doi.org/10.15585/mmwr.ss6503a1.
14. Manfred Del Fabro, Bernd Münzer, and Laszlo Böszörmenyi. 2013. *Smart Video Browsing with Augmented Navigation Bars*. Springer Berlin Heidelberg, Berlin, Heidelberg, 88–98. DOI: http://dx.doi.org/10.1007/978-3-642-35728-2_9
15. Philippe Dreuw and Hermann Ney. 2008. Towards automatic sign language annotation for the elan tool. In *Proceedings of LREC*.
16. Damien Dupre, Daniel Akpan, Elena Elias, Jean-Michel Adam, Brigitte Meillon, Nicolas Bonnefond, Michel Dubois, and Anna Tcherkassof. 2015. Oudjat: A configurable and usable annotation tool for the study of facial expressions of emotion. *International Journal of Human-Computer Studies* 83 (2015), 51 – 61. DOI: http://dx.doi.org/10.1016/j.ijhcs.2015.05.010

17. Laura Hänninen and Matti Pastell. 2009. CowLog: Open-source software for coding behaviors from digital video. *Behavior Research Methods* 41, 2 (2009), 472–476.

18. Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 6536–6546. DOI: http://dx.doi.org/10.1145/3025453.3025821

19. Ingrid Maria Hopkins, Michael W. Gower, Trista A. Perez, Dana S. Smith, Franklin R. Amthor, F. Casey Wimsatt, and Fred J. Biasini. 2011. Avatar Assistant: Improving Social Skills in Students with an ASD Through a Computer-Based Intervention. *Journal of Autism and Developmental Disorders* 41, 11 (01 Nov 2011), 1543–1555. DOI: http://dx.doi.org/10.1007/s10803-011-1179-z

20. Debora M Kagohara, Larah van der Meer, Sathiyaprakash Ramdoss, Mark F OâĂŹReilly, Giulio E Lancioni, Tonya N Davis, Mandy Rispoli, Russell Lang, Peter B Marschik, Dean Sutherland, and others. 2013. Using iPods and iPads in teaching programs for individuals with developmental disabilities: A systematic review. *Research in developmental disabilities* 34, 1 (2013), 147–156.

21. Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. 2008. DRAGON: A Direct Manipulation Interface for Frame-accurate In-scene Video Navigation. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 247–250. DOI: http://dx.doi.org/10.1145/1357054.1357097

22. Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. 2012. DragLocks: Handling Temporal Ambiguities in Direct Manipulation Video Navigation. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 623–626. DOI: http://dx.doi.org/10.1145/2207676.2207764

23. Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. 2016. Discovering Objects of Joint Attention via First-Person Sensing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

24. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven Interaction Techniques for Improving Navigation of Educational Videos. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM, New York, NY, USA, 563–572. DOI: http://dx.doi.org/10.1145/2642918.2647389

25. E. I. Konstantinidis, A. Luneski, C. A. Frantzidis, P. Costas, and P. D. Bamidis. 2009. A proposed framework of an interactive semi-virtual environment for enhanced education of children with autism spectrum disorders. In *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*. 1–6. DOI: http://dx.doi.org/10.1109/CBMS.2009.5255414

26. Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2176–2184.

27. Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM, New York, NY, USA, 541–550. DOI: http://dx.doi.org/10.1145/2642918.2647366

28. Soichiro Matsuda, Eleuda Nunez, Masakazu Hirokawa, Junichi Yamamoto, and Kenji Suzuki. 2017. Facilitating Social Play for Children with PDDs: Effects of Paired Robotic Devices. *Frontiers in Psychology* 8 (2017), 1029. DOI:http://dx.doi.org/10.3389/fpsyg.2017.01029

29. David McNaughton and Janice Light. 2013. The iPad and Mobile Technology Revolution: Benefits and Challenges for Individuals who require Augmentative and Alternative Communication. *Augmentative and Alternative Communication* 29, 2 (2013), 107–116. DOI: http://dx.doi.org/10.3109/07434618.2013.784930 PMID: 23705813.

30. Peter Mundy, Christine Delgado, Jessica Block, Meg Venezia, Anne Hogan, and Jeffrey Seibert. 2003. Early social communication scales (ESCS).

31. Leslie Neely, Mandy Rispoli, Siglia Camargo, Heather Davis, and Margot Boles. 2013. The effect of instructional use of an iPad on challenging behavior and academic engagement for two students with autism. *Research in Autism Spectrum Disorders* 7, 4 (2013), 509 – 516. DOI:http://dx.doi.org/10.1016/j.rasd.2012.12.004

32. Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2013. Direct Manipulation Video Navigation in 3D. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 1169–1172. DOI: http://dx.doi.org/10.1145/2470654.2466150

33. Maria B Ospina, Jennifer Krebs Seida, Brenda Clark, Mohammad Karkhaneh, Lisa Hartling, Lisa Tjosvold, Ben Vandermeer, and Veronica Smith. 2008. Behavioural and developmental interventions for autism spectrum disorder: a clinical systematic review. *PloS one* 3, 11 (2008), e3755.

34. Suporn Pongnumkul, Jue Wang, Gonzalo Ramos, and Michael Cohen. 2010. Content-aware Dynamic Timeline for Video Browsing. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM, New York, NY, USA, 139–142. DOI: http://dx.doi.org/10.1145/1866029.1866053

35. Yael Pritch, Alex Rav-Acha, Avital Gutman, and Shmuel Peleg. 2007. Webcam Synopsis: Peeking Around the World. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Washington, DC, USA, 1–8. DOI: http://dx.doi.org/10.1109/ICCV.2007.4408934

36. Alex Rav-Acha, Yael Pritch, and Shmuel Peleg. 2006. Making a Long Video Short: Dynamic Video Synopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Washington, DC, USA, 435–441. DOI: http://dx.doi.org/10.1109/CVPR.2006.179

37. Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. 2013. Gaze Locking: Passive Eye Contact Detection for Human-object Interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. 271–280. DOI: http://dx.doi.org/10.1145/2501988.2501994

38. Kenji Suzuki, Taku Hachisu, and Kazuki Iida. 2016. EnhancedTouch: A Smart Bracelet for Enhancing Human-Human Physical Touch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. 1282–1293. DOI: http://dx.doi.org/10.1145/2858036.2858439

39. B. Takacs. 2005. Special education and rehabilitation: teaching and healing with interactive graphics. *IEEE Computer Graphics and Applications* 25, 5 (Sept 2005), 40–48. DOI:http://dx.doi.org/10.1109/MCG.2005.113

40. Joshua Wade, Arpan Sarkar, Amy Swanson, Amy Weitlauf, Zachary Warren, and Nilanjan Sarkar. 2017. Process Measures of Dyadic Collaborative Interaction for Social Skills Intervention in Individuals with Autism Spectrum Disorders. *ACM Trans. Access. Comput.* 10, 4, Article 13 (Aug. 2017), 19 pages. DOI: http://dx.doi.org/10.1145/3107925

41. Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. *A Discriminative Feature Learning Approach for Deep Face Recognition*. Springer International Publishing, Cham, 499–515. DOI: http://dx.doi.org/10.1007/978-3-319-46478-7_31

42. Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of LREC*, Vol. 2006. 5th.

43. Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D. Abowd, and James M. Rehg. 2012. Detecting Eye Contact Using Wearable Eye-tracking Glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 699–704. DOI: http://dx.doi.org/10.1145/2370216.2370368

44. Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg. 2015. Detecting bids for eye contact using a wearable camera. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–8. DOI: http://dx.doi.org/10.1109/FG.2015.7163095

45. Hiroshi Kera Ryo Yonetani Keita Higuchi Yifei Huang, Minjie Cai and Yoichi Sato. 2017. emporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Video. In *The International Conference on Computer Vision (ICCV) Workshops*.

46. Hanan Makki Zakari, Minhua Ma, and David Simmons. 2014. *A Review of Serious Games for Children with Autism Spectrum Disorders (ASD)*. Springer International Publishing, Cham, 93–106. DOI: http://dx.doi.org/10.1007/978-3-319-11623-5_9

47. Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery. DOI: http://dx.doi.org/10.1145/3126594.3126614

48. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520.

49. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017-05-18). https://perceptual.mpi-inf.mpg.de/files/2017/05/zhang_cvprw2017.pdf