A Scalable Off-the-Shelf Framework for Measuring Patterns of Attention in Young Children and its Application in Autism Spectrum Disorder

Matthieu Bovery, Geraldine Dawson, Jordan Hashemi, *Member, IEEE* and Guillermo Sapiro, *Fellow, IEEE*

Abstract—Autism spectrum disorder (ASD) is associated with deficits in the processing of social information and difficulties in social interaction, and individuals with ASD exhibit atypical attention and gaze. Traditionally, gaze studies have relied upon precise and constrained means of monitoring attention using expensive equipment in laboratories. In this work we develop a low-cost off-the-shelf alternative for measuring attention that can be used in natural settings. The head and iris positions of 104 16-31 months children, an age range appropriate for ASD screening and diagnosis, 22 of them diagnosed with ASD, were recorded using the front facing camera in an iPad while they watched on the device screen a movie displaying dynamic stimuli, social stimuli on the left and non-social stimuli on the right. The head and iris position were then automatically analyzed via computer vision algorithms to detect the direction of attention. We validate the proposed framework and computational tool showing that children in the ASD group paid less attention to the movie, showed less attention to the social as compared to the non-social stimuli, and often fixated their attention to one side of the screen. These results are expected from the ASD literature, here obtained with significantly simpler and less expensive attention tracking methods. The proposed method provides a low-cost means of monitoring attention to properly designed stimuli, demonstrating that the integration of stimuli design and automatic response analysis results in the opportunity to use off-the-shelf cameras to assess behavioral biomarkers.

Index Terms—Autism spectrum disorder, Computer vision, Gaze-tracking, Attention, Off-the-shelf cameras, Stimuli design

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by qualitative impairments in social interaction and the presence of restricted and repetitive behavior [1]. Studies of children in the first three years of life have shown that a failure to orient and lack of attentional preference for social information distinguishes children with ASD from those with typical development and other developmental delays [2], [3]. These atypical patterns of social attention are manifested early in life [4], [5], [6], and while not exclusive to ASD, are known to be strong candidates for ASD and developmental disorders biomarkers, even genetically influenced [7]. Thus, the development of feasible and reliable methods for assessing early-emerging differences in patterns of attention is of significant interest, with the goal of eventually developing scalable behavioral analysis tools for early screening, diagnosis, and treatment monitoring.

To further our understanding of the differences in social processing in children with ASD, researchers have utilized eye-gaze tracking to measure gaze responses to dynamic visual stimuli. Such measures have been shown to differentiate ASD from other populations starting at the age of 6 months [8], [9], [10], [11], [12]. It has been demonstrated that children with ASD show differential gaze patterns compared to typically developing children, characterized by a lack of attention preference for socially salient stimuli [13], [14]. Other studies have shown that children with autism are less likely to shift their attention throughout the stimuli and explore scenes containing both social and non-social components. [15]. These studies have used either expensive eye-tracking devices or advanced methods, such as dark pupil-corneal reflection video-oculography techniques. Such sensing and acquisition approaches are not scalable and are not readily applicable in natural environments. Furthermore, such studies tend to use a region-of-interest based approach for analysis in which the feature of interest is the amount of time fixating at a specific region of interest. This approach also often fails to capture the dynamic quality of attention, including important temporal patterns such as how attention shifts in response to the stimulus dynamics.

In the current work we present a framework that aims at confirming these previous results but with a significantly simpler and less expensive attention tracking method (related approaches are discussed later in the manuscript). A dynamic movie that contained salient social and non-social stimuli was used to investigate attention patterns. The capability of sensing and analysis tools, namely an off-the-shelf video camera and computer vision, were taken into

- *M. Bovery is with the EEA Department, ENS Paris-Saclay, Cachan, FRANCE. He performed this work while visiting Duke University. E-mail: matthieu.bovery@ens-cachan.fr*
- *J. Hashemi and G. Sapiro are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. G. Sapiro is also with BME, CS, and Math at Duke University.*
- *G. Dawson is with the Department of Psychiatry and Behavioral Sciences, Duke Center for Autism and Brain Development, and the Duke Institute for Brain Sciences, Durham, NC.*

Manuscript received ?; revised ?.

consideration for designing that movie. The movie screen displayed a social stimulus that was looking toward and interacting with the viewer on the left side of the screen, and a non-social but visually interesting moving object on the right. This required only a right versus left attention discrimination to evaluate dynamic preference for social or non-social stimulus. Figure 2 shows screen shots of the designed movie. We presented this movie on an iPad to young children with ASD and to typically-developing children, and used the front camera in the iPad to film their responses. After automatically processing the recorded videos, we used to position of the pupils and the head angles to determine which part of the screen they were looking at.

The first hypothesis we want to confirm within this framework is that children with ASD would exhibit a reduced amount of attention to the movie overall, as found in previous studies. This was tested by comparing the overall amount of time spent looking at the movie (regardless of side) by children with ASD vs non-ASD children. The second hypothesis used for validation of the proposed framework was that ASD children would exhibit an attentional preference for the non-social stimulus (dynamic toy) as compared to the social stimulus (woman singing nursery rhymes while making eye contact), here tested once again by automatically computing the attention direction from the recorded video. The last hypothesis we tested was that children with ASD are more likely to fixate on one side of the screen, regardless of stimulus. For this we split the movie in time segments, with stimuli changing on the left or right (see Figure 2 in the Methods section), and then analyzed the attention for each one of these time segments.

Using the carefully designed stimuli with the standard RGB camera and simple computer vision algorithms to validate these predictions, we demonstrated that scalable tools can be used to measure the same type of eye gaze biomarkers that previously necessitated high-end eye-tracking devices. It is critical to note that contrary with the standard in the literature, where available video stimuli are used, here we stress the need to integrate the stimuli design with the available device (RGB consumer camera on a tablet in this case), task (distinguish between social an non-social for example), and algorithm design and robustness capabilities (region vs pixel accuracy for example). Therefore, while available databases, e.g., [16], can and should be used for algorithm validation in some cases (e.g., validating affect computation), they become less appropriate for new tasks and the integrated approach here pursued.

2 METHODS

2.1 Participants

Participants in this study were 104 toddlers between 16 and 31 months of age. This is the age range at which gold standard diagnostic methods have been validated. Twenty-two children were diagnosed with autism spectrum disorder (ASD). Diagnosis was based on both expert clinical judgment by a licensed clinical psychologist with expertise in ASD and the Autism Diagnostic Observation Scale-Toddler Module [17], which can be used with toddlers as young as 12 months of age (mean age =26.19 months, standard deviation σ=4.07 months). The remaining 82 toddlers were typically-developing (non-ASD) or had developmental delay (mean age of M=21.9; σ=3.78 months). Participants were recruited either from primary care pediatric clinics at Duke Health, directly by a research assistant or via referral from their physician, or by community advertisement. For the clinic's recruitment, a research assistant approached participants at their 18- or 24-month well child visit, when all children in the clinic are screened for ASD with the Modified Checklist for Toddlers-Revised with Follow-up Questions (M-CHAT-R/F) [18]. Toddlers with known vision or hearing deficits were excluded. Toddlers were also excluded if they did not hear any English at home or if parents/guardians did not speak and read English sufficiently for informed consent. All parents/legal guardians of participants gave written, informed consent, and the study's protocol was approved by the Duke University School of Medicine Institutional Review Board.

The majority of children recruited into the study had already received screening with a digital version of the M-CHAT-R/F as part of a quality improvement study in the clinic [19]. Participants from community recruitment received ASD screening with the digital M-CHAT-R/F prior to the tablet assessment reported in this work [20]. As part of their participation in the study, children who failed the M-CHAT-R/F or for whom parents/legal guardians or their physicians had concerns about possible ASD, underwent diagnostic and cognitive testing with a licensed psychologist or trained research-reliable examiner overseen by a licensed psychologist. Testing consisted of Autism Diagnostic Observation Schedule Toddler Module (ADOS-T) and Mullen Scales of Early Learning (MSEL) [21], [22]. Children who received a diagnosis of ASD based on the ADOS-T and clinician assessment were referred to early intervention services.

Children were enrolled consecutively and screened for ASD, resulting as expected in a greater number of typical children compared to those with ASD (there is a 1:59 prevalence of ASD in the US). For the goal of the work here presented, namely introducing a computational integrated stimulus-device-algorithm design for scalable attention analysis (here illustrated for ASD), this unbalance is not a concern. The size of each class is sufficient to illustrate the virtue of the proposed approach and to provide initial findings, to be fully statistically validated in subsequent studies (e.g., [23]).

2.2 Stimulus and measures

During a regular clinic visit (no special setup, just a standard pediatric clinic room), we asked the participants to sit on a caregiver's lap while watching a movie on a tablet (iPad 4th generation) [20]. Since we monitor the movement and position of the head as detailed below, seating on a lap, as it is common for protocols at this age (e.g., [24]), was not found to be a problem but can be considered a factor to be improved in the future. The tablet was placed on a stand approximately 3 feet away from the child to prevent her/him from touching the screen; see Figure 1. The brief movie displayed in landscape mode and split in two regions: on the left side a woman is singing to the child, and on the right side a moving toy making some noise to also

try to draw the participant's attention. The woman as well as the toy changed throughout the movie; see Figure 2. The entire movie was one minute. Parents were asked to attempt to keep the child seated on their lap, but to allow the child to get off their lap if they became too distressed to stay seated. The iPad's front facing camera recorded the child's face, at 1280×720 and 30 frames per second resolution, while they were watching the movie. This comprised all of the sensed data used by the automatic computer vision algorithm to measure attention.

The stimuli here used, Figure 2, are common in the ASD literature to represent social and non-social stimuli, e.g., [9], [24], [25], [26]. The social and non-social halves differ also in color and dynamics, and one may argue that this might influence the child's attention as well (and not just the social or non-social aspects of the halves). This influence, even if it exists, is not affecting the computational approach here introduced, since the goal is to detect the direction the participant is looking at, regardless (at this stage) of the reason they are looking at it, and this is accomplished by the proposed algorithm described next. Moreover, regardless of the exact reason for the left-right attention preference, there is still a fundamental difference between ASD and non-ASD groups, as we will show in subsequent sections, providing potential value as a behavioral biomarker, for example for screening.

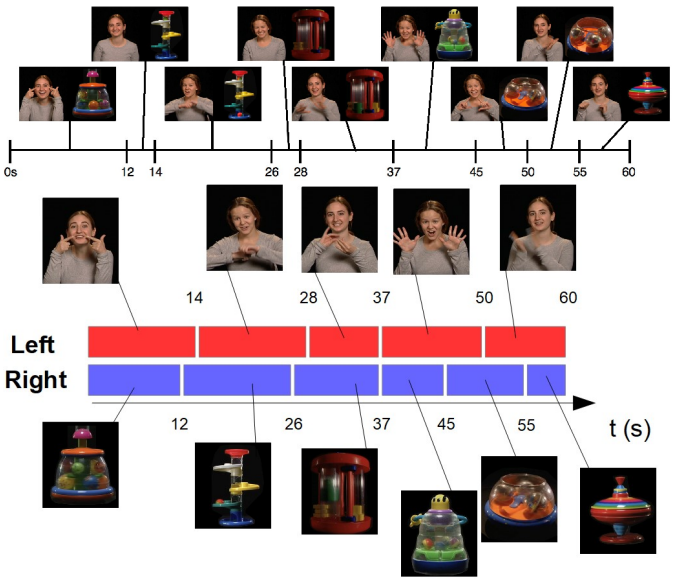

Fig. 2. Screen shots of the designed stimulus. The movie showed a social stimulus on the left (singing women) and a non-social on the right (toys), top figure. Both halves changed during the 60 seconds, as further detailed in the bottom figure, defining a total of 9 temporal blocks. The movie was carefully designed in an integrated fashion, considering not only the type of stimulus but also the used sensing device (a regular camera) and capabilities of the automatic analysis algorithm.

2.3 Head position tracking

The children's responses were recorded by the frontal camera of the tablet at 1280 by 720 resolution and 30 frames per second; see Figure 1. We used the computer vision algorithm (CVA) detailed in [20] to automatically detect and track 51 facial landmarks on every child's face, allowing for detection of head, mouth, and eye position [27]. These landmarks are used here and for subsequent steps of the proposed computational algorithm, and follow extensive work and validation, see [20], [27], [28]. Algorithms based on region and not pixel accuracy, as here proposed when integrated with properly designed stimuli, provide further robustness (accuracy needs to be region based on not pixel based). Any further improvement in the landmarks detection (see for example [29], [30]) can immediately be incorporated into the proposed framework since these are the inputs to our algorithms.

We estimated the head positions relative to the camera by computing the optimal rotation parameters between the detected landmarks and a 3D canonical face model [31].

2.4 Direction of attention tracking

We implemented an automatic method to track frame-by-frame the direction of the child's attention from the data mentioned above. We also took into account the fact that the child might not be attending to the movie at all (see also [29], [30], [32], [33], [34] for alternative approaches to detect if the participant is attending the stimulus).

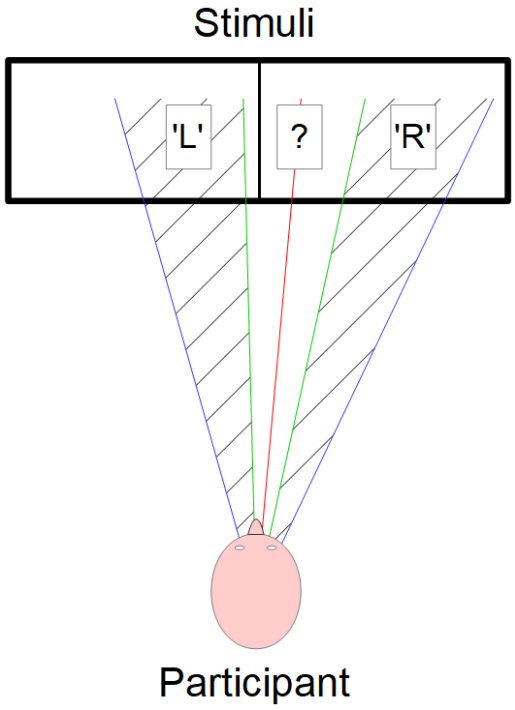

Fig. 3. Illustration of the first component of the attention tracking method. We use the extreme yaw angles values (in blue) to determine the midrange yaw angle value (in red). Then, we define two thresholds (in green) by adding or subtracting 10% of the difference between the midrange value and the extreme values to the midrange value. With those thresholds, we determine wheter the partcipant is looking at the left part of the stimuli ('L'), at the right part of the stimuli ('R'), or if the yaw angle value was not large enough to conclude ('?'). In this last case, we used the landmarks to make a decision (see Figure 4 and text for more details).

For detecting the direction of attention, we first used the value of the yaw angle obtained from the head position as described in the previous section (see Figure 3). For a given

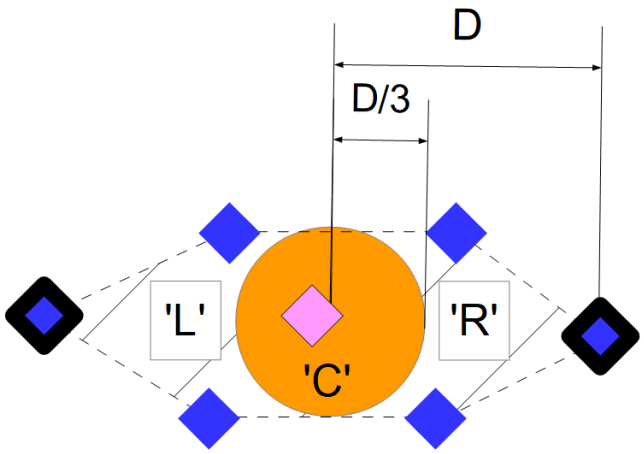

Fig. 4. Zoom on the eye's landmarks (see Figure 1 for more details) to illustrate the second component of the attention tracking method. We use the landmarks at the edges of the eye (bold landmarks) to estimate the position of the middle of the eye and the distance between this middle and both edges (D). Then, we check wheter or not the pupil's landmark (pink landmark) is close enough to one of the edges to conclude the attention direction. We considered it 'close enough' if the distance between the pupil's landmark and the center of the eye is greater than D/3. If not, we assume that the participant is looking somewhere in the middle of the stimuli ('C'), as is the case in this example. We use this method on both eyes.

subject, we considered the midrange value of the yaw angle among all the frames in order to take into account the initial position of the head. Then we compared the difference between the yaw angle of each frame and the midrange value to the difference between the most extreme value for the same frame and the midrange value. If the difference between the yaw angle and the midrange value was at least 10% larger than the difference between the midrange value and the extreme value, then this was considered large enough for detection of attention direction (provided that it is not too large to indicate no-attention, see [20]), and we could easily conclude whether the child was looking at the left or the right side of the screen. If not, we exploited the landmarks (see Figure 1 and Figure 4). In this case, we compared the position of the iris to the edges of the eyes. We looked in particular at whether or not the distance between the iris and the center of the eye was larger than 1/3 of the distance between the middle and either edge of the same eye. If both irises were close enough, according to this criterion, to the same edge of their respective eyes, then we once again assessed the direction of the attention (gaze). The results were found to be robust to the selection of 10% and 1/3, and these values can be modified if more or less robust measurements are desired. Note that these measurements are relative and per-child, adding robustness across participants and to the distance to the screen.

Using this method we were able to label most of the frames either as 'L' (for attending to the left) or 'R' (for attending to the right). In some frames the computer vision algorithm failed to properly track the landmarks due to the child not facing the camera. The algorithm would then output non-numerical data and we labeled those frames with the standard 'NaN' (for 'Not a Number'). In addition, in some cases, neither the value of the yaw angle nor the positions of the irises within the eyes were sufficient to conclude the direction of attention. We then simply assumed that the child was looking somewhere in the middle of the screen and labeled those frames 'C' (for 'Center'). We could have also ignored these frames, since overall at 30 frames per second and one minute of recording, we had ample data to work for analyses. Indeed, within the frames where a participant was paying attention (frames labelled either 'L', 'R,' or 'C'), only about 0.5% of them are labelled 'C.' Overall, 90.8% of the study frames were labeled 'L' or 'R.' It is important once again to stress that with the joint-design of stimulus, sensing, and analysis, the fact that we were going to have inconclusive frame labeling was taken into account.

2.5 Temporal block analysis

In addition to measuring attention and direction of attention, we also studied fixation and the attention responses to changes in the stimuli. As mentioned before, both the woman (social stimulus) and the toy (non-social stimulus) are changing multiple times throughout the video, Figure 2. In other words, there are several women each with a different rhyme and several toys each with different characteristics. As a proxy to fixation, we tracked the participant's attention shifts to those changes in stimulus. Hence, we split our data into temporal blocks such that we distinguish when there was a change in either the social or the non-social stimulus (both do not always change simultaneously). In other words, the boundaries of each temporal block were given by a dynamic change of the toy (non-social), the woman (social), or both. With this approach, we created nine time blocks of different lengths (see captions in Figure 2), over which we integrated the previously described per-frame results (based on a simple majority). We therefore obtained for each participant nine labels, one per temporal block, categorized as 'L', 'R', 'C,' and 'NaN'. We used this to examine whether the child's attention shifted when to the novel stimulus when there was a change.

We should note that we also experimented merging the short intervals with the consecutive longer ones, obtaining a total of 6 temporal blocks of approximately equal length. Since the same qualitative results were obtained, we kept the original 9 temporal blocks to incorporate the short transitions as well in the analysis.

3 Results

The overall difference of attention between the ASD and the control groups is considered first, shown in Figure 5. We defined attention frames as the frames labeled either 'L,' 'R,' or 'C.'

For the ASD group, the mean value was M=1,406 frames, and the standard deviation σ=460.3 frames. In comparison, M=1,717 frames and σ=228.3 frames for the control group. The number of participants who were paying attention to fewer than 1,000 frames is 18.2% for the ASD group, whereas it was only 1% for the control group. About 74.4% of the control participants were paying attention to the whole movie, while about 68.2% of the participants with ASD were not attending at some point of the movie.

Next, we studied the attentional preferences of each participant by dividing the screen into two halves (social versus

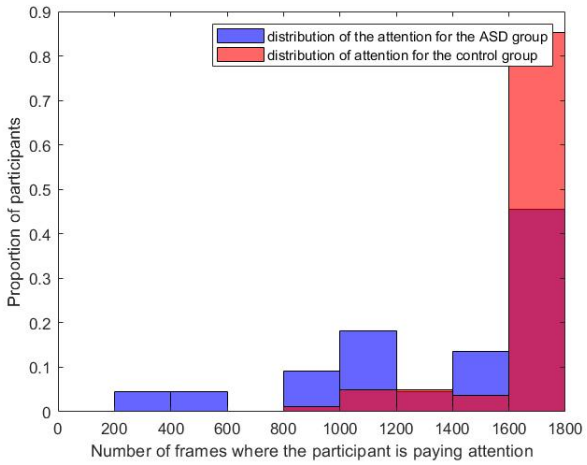

Fig. 5. Proportion (vertical axis) of ASD (in blue) and control (in red) participants paying attention to the total number of movie frames indicated in the horizontal axes.

non-social). As illustrated in Figure 6, we examined the proportion (%) of frames during which the participant was looking right (non-social), as a function of the proportion (%) of frames during which the participant was looking left (social stimulus). Those proportions were calculated by dividing the number of frames during which the participant was looking at the given stimulus by the total amount of frames during which the child was paying attention.

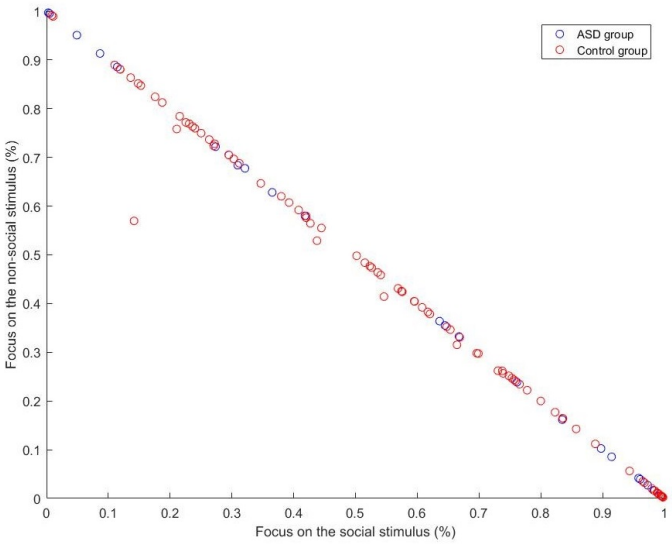

Fig. 6. We plot the ratio between social and non-social attention, each participant being a circle, with ASD group in blue and control group in red.

The pattern shown in Figure 6 suggests that children with ASD and non-ASD children were attending to the movie in very similar ways. The means and standard deviations for attention to social stimulus were M=52%, σ=35% for the ASD group and M=55%, σ=29% for the control group. For the non-social stimulus, results were M =48%, σ=35% for the ASD group and M=44%, σ=29% for the control group. However, when we examined the extreme values, an interesting pattern emerged, revealing a feature that distinguished ASD from non-ASD children. First, the proportion of participants who paid attention to the social stimulus for greater than 95% of frames was similar across groups, 18% for the ASD group and 15% for the control group. In contrast, the proportion of participants who paid attention to the non-social stimulus for greater than 90% of frames was 18% for the ASD group compared to only 2% for the control group, indicating that it is very rare for non-ASD participants to spend most of their attention time on the non-social stimulus. Some points in Figure 6 are not on the diagonal, indicating that those participants are looking at the center of the stimuli for a significant number of frames. Almost 95% of the children devoted less than 1% of their attention to the center of the stimuli. Out of the 5% that did not, all were within the control group.

Next, we studied the temporal pattern of attention direction taking into account the temporal block data, i.e., changes in the stimuli. We computed two 3D-histograms (one for each group, ASD and control) reflecting the proportion of the attention toward either side of the screen (Figure 7). Each value in the histogram position (i,j) (i,j=1..9) represents the percent of participants in the group that spent i temporal blocks attending to the left and j blocks attending to the right.

Examining the control group, we can see that about 60% of the points are on the diagonal (points that add to 9, the total number of temporal blocks), which means those non-ASD children have their nine blocks labeled either 'L' or 'R.' Alongside the diagonal, the points are uniformly distributed, if not for two spikes. The one on right corresponds to the 15.8% participants that have all their blocks labeled 'L.' The other one in the center corresponds to the 11% of the participants that have 4 blocks labeled 'L' and 5 blocks labeled 'R.' The mean value for the number of temporal blocks spent looking at the social stimuli is M=4.7 blocks and the standard deviation σ=2.8 blocks. For the number of blocks spent looking at the non-social stimuli, M=3.2 blocks and σ=2.7 blocks.

For the ASD group, only 28% of the points are located on the diagonal (meaning only 28% are fully attending). More than 36.4% of the participants have at least 8 out of their 9 blocks labeled either 'L' or 'R,' and 77% of them have less than two blocks labeled 'R' or less than two blocks labeled 'L'. Moreover, 59% of the children with ASD have less than one block labeled 'R' or less than one block labeled 'L.' All these results indicated a very one-sided attention orientation. The mean number of blocks spent looking at the social stimulus was M=3.3 blocks and the standard deviation σ=3.2 blocks. The mean number of blocks spent looking at the non-social stimulus was M=3.1 blocks and σ =3.3 blocks.

We now illustrate how the proposed computational tool opens the door to further granularity, investigating the actual dynamic pattern of attention when the stimulus changes, see Figure 8. As we see from Figure 2, the left/social part of the stimulus changes 4 times (intervals 2-3, 4-5, 5-6, 7-8), while the right/non-social changes 5 times (intervals 1-2, 3-4, 5-6, 6-7, 8-9); these are indicated in the horizontal axis of each one of the subfigures in Figure 8. We then compute how the attention switches when such stimulus changes occur. For illustration of the details in the

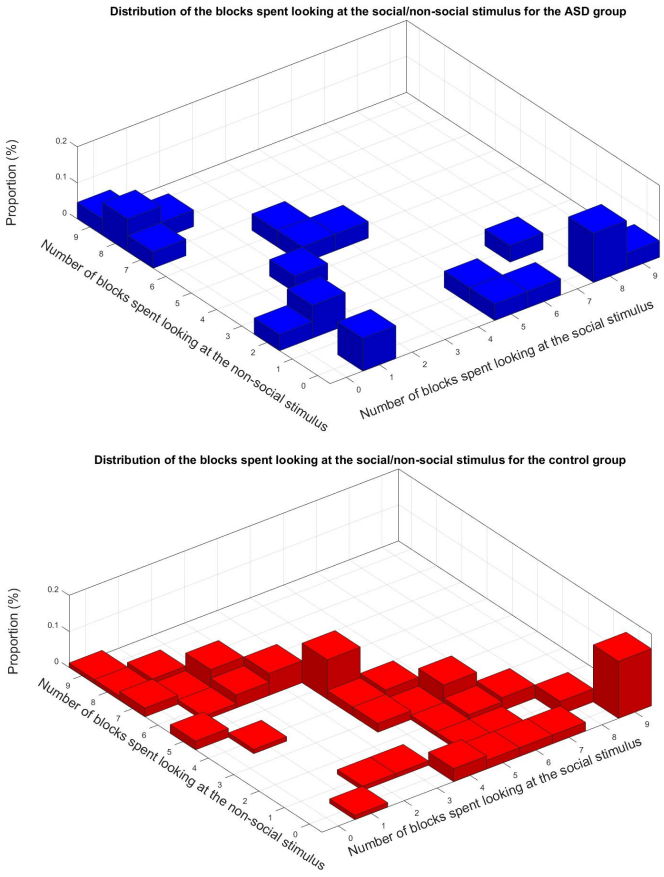

Fig. 7. Histograms of temporal attention direction for the 9 different temporal blocks resulting from the stimulus changing (see Figure 2). The 3D histograms indicate how the different participants spent their attention for each one of the 9 time blocks in the stimulus, meaning each entry (i,j) represents the proportion of participants spending i blocks attending to the left and j blocks attending to the right

sub-figures, we will assume as a running example that the change is happening on the left part of the stimulus. The first subfigure (a) shows the percent of participants that shifted their attention toward the side where the change is happening, normalized by all the participants. Considering the running example, this is (participants that were looking R and then looked L)/(all participants). This is repeated in the next subfigure (b), now normalized by the participants that were looking in the direction opposite to where the change happened. In the running example, this is (participant that were looking R but then looked L)/(participants that were looking R). The third subfigure (c) shows the percent of participants that were looking where the change happened and that shifted their attention away from it, normalized by the participants that were looking where the change happened. In the running example, this is (participant that were looking L but then looked R)/(participants that were looking L). Finally, the last subfigure (d) shows the percent of participants that shifted their attention to where the change happened, but then shifted their attention back away from it, normalized by the participants that shifted their attention to the side where the change happened. In the running example, this is (participant that were looking R and then looked L and then looked R again)/(participants that were looking R and then looked L). While the total number per class/stimulus switch is relatively small (indicated by the numbers in each bar) to perform full statistical analysis, we start seeing an interesting pattern depending on what side, left/social or right/non-social changed. More importantly, this example further illustrates that the tool here developed can provide information about granular and dynamic shifts of attention, all with an off-the-shelf camera and an algorithm tuned to the presented and carefully designed active sensing.

4 Discussion

As mentioned before, the goal of this work is to derive a computational framework for attention monitoring and then validate it confirming previous results, this time with a significantly simpler and less expensive method, thereby providing a low-cost scalable tool and paradigm to measure attention. We now discuss how the results in the previous section support this and at the same time provide potential directions to investigate the proposed technique and findings in larger studies.

4.1 Deficit in overall attention

We first evaluated whether children with ASD differ from non-ASD children in terms of their overall attention to the presented movie. We automatically computed, frame by frame, whether or not the participant was looking at the iPad screen, and then we compared the number of frames during which the child was looking at the screen across the two groups (Figure 5). We confirmed the hypothesis that children with ASD exhibited reduced attention to the movie overall. This was further supported with the block analysis (Figure 7), where we can see that the density of points close to the origin is significantly higher for the ASD group than it is for the control group. Those points are indicating that the child had most of their blocks labeled 'NaN,' which means that the child was not attending to the screen over multiple periods of time.

These results demonstrate the usefulness of low cost ubiquitous devices, consumer cameras available on tables or mobile phones, to measure attention. This method is in sharp contrast with the high-end and expensive hardware that is common in most ASD studies. Secondly, these results can be informative as one feature that could contribute to an algorithm/scoring for ASD screening. For example, we could consider that a participant paying attention to less than a certain percentage of frames would be one feature more commonly associated with ASD. For our data, for example, considering 1,000 frames, the values of the precision, recall and F1-score are P=0.8, R=1, and F1=0.89, respectively. These results are only a first step, and their statistical power needs to be investigated with larger populations. In addition, lack of attention is not an ASD exclusive behavior, and as such it should be considered as one of many scores in a full evaluation, similarly to the current standard of care which includes the observation of multiple behaviors.

4.2 Attention to social versus non-social stimuli

Within our scalable framework, we also tested whether the ASD group attended more to the non-social than the social

stimulus as compared to the control group. We tracked attention on a frame-by-frame basis. We then examined the proportion of frames spent looking at the right, where the non-social part of the stimulus was displayed, versus the proportion of frames spent looking at the left, where the social part of the stimulus was displayed (figures 1 and 4).

Our first analyses simply comparing the average number of frames looking at the social versus non-social stimuli did not yield group differences. However, our analyses could be further improved by splitting the stimulus regions, e.g., in 4 sub-regions instead of just 2, and looking within the social stimulus to test if the ASD participants are less likely to look at the face of the woman (top part of the left side), as suggested by earlier studies [35], [36], [8]. Our preliminary work indicates that we can further obtain such increased accuracy with the type of sensors and computer vision tools here considered.

Looking at the extreme values with respect to how attention was distributed across the social and non-social stimuli revealed compelling results. We observed that when a participant with ASD paid the majority of their attention to only one side of the screen, it was equally likely to be toward the social or non-social side. On the other hand, if a control participant exhibited the same behavior of attending mostly one side of the screen, it was seven times more likely that the child was looking at the side of the screen displaying the social stimulus. This feature could also potentially be used as an additional risk marker for ASD as part of an off-the-shelf device and a simple test. As discussed in the previous section, the statistical power of this measurement as a risk factor for ASD deserves future study in large populations.

Finally, these results and data also showed that a very high percent of participants with ASD focus almost solely on a single side of the screen and were less likely to switch their attentional focus from one side to the other. This aspect of fixation is further discussed next.

4.3 Attention shifting

We also used this recording and analysis framework to study fixation behavior, that is, the degree to which the child shifts their attention from right to left throughout the movie. We split the data into temporal blocks corresponding to different social or non-social stimuli, Figure 2. We then determined the most popular label over each temporal block and computed the corresponding per-block frequencies (Figure 7). We are here looking at the participants that are paying attention to most of the stimulus, that is, the points that are close to the diagonal in the 3D histograms.

We can clearly distinguish different patterns between the ASD and the control groups. The non-ASD children followed two main patterns: while some of the children spent most of the time attending the social stimulus, most distributed their attention between both the social and the non-social ones. The vast majority of the children with ASD, on the other hand, attended almost solely at either the left or the right part of the screen, supporting the previous conclusions and further demonstrating we can use this framework to understand attention shifting. Future work should include also switching the social/non-social stimuli to be displayed on both sides of the screen to understand more fully what evokes attention shifts. This is partially discussed next.

Finally, we demonstrated how to use the developed tool to carefully study the attention dynamics, and studied the patterns of attention shift as a result of stimulus changes, Figure 8. While the actual population is relatively small, differences are starting to appear depending on the stimulus region that is actually changing (social or non-social).

5 Conclusions

By replicating the type of attention patterns for children with ASD previously reported in the literature, we validated the possibility of using scalable tools to measure this important biomarker. Contrary to the common use of high-end devices to measure attention, this work exploited ordinary cameras available in tablets and mobile phones, integrating from the start the stimulus design with the sensing and analysis. The statistical power of the reported results, initially as a screening tool, need to be investigated with larger populations, part of our ongoing activity at the NIH Autism Center of Excellence [23], where these techniques will be tested in thousands of participants.

We validated our framework based on three hypotheses previously derived from studies using state-of-the-art eye-tracking technology. First, we monitored the attention of both the ASD and the control group and showed that the ASD participants were more likely to have reduced attention to the movie overall. We next examined differences between social and non-social attention. We found that, whereas it was very rare for a child without ASD to focus the majority of their attention on the non-social stimuli, this occurred much more often among the children with ASD. Thus, this biomarker has potential strong sensitivity as a risk marker for ASD. Finally, we took into account the temporal changes in the stimulus to investigate patterns of fixation and shifting of attention. We showed that participants with ASD are more likely to fixate on either part of the movies (stimulus social/non-social regions) than the non-ASD children, providing yet an additional potential biomarker.

While the work here presented concentrated on ASD, using stimuli and validation paradigms from the ASD research literature, there is extensive literature supporting the fact that attention as a biomarker is critical in a wide range of neuropsychiatric conditions beyond ASD, such as attention deficit hyperactivity disorder (ADHD) and anxiety. The direction of attention, and not just attention to the stimulus itself, can also be of use for intervention, e.g., [32], [33]. Furthermore, this framework here presented can be integrated with robots as described in [34]. Note that contrary with the tool exploited there, namely [29], here we co-design stimulus and computation. Our initial experience, e.g., [28], [37], indicates that such active sensing and integrated approach is more robust and engaging, over 85% of usable frames vs. only about 50% reported in [34] (although for different environments and protocols).

While attention is a very important behavioral feature in ASD screening, diagnosis, and symptoms monitoring, it should be considered together with other measurements, from body posture and affect to audio. Each different behavior will provide information in the complex and diverse

structure of ASD, and all should be sensed with scalable and engaging protocols, e.g., [28], [34], [37], [38].

The stimulus and paradigms used in this work relied on measurements of attention to the right or left side of the screen. Our initial work [39], [40] indicates that we are able to further split the screen into more sub-regions, allowing for greater flexibility in stimuli design and measurement of behavioral biomarkers. Regardless, as here demonstrated, in order to achieve true scalability, we must use off-the-shelf sensors and for that, stimuli design needs to be integrated with sensing and analysis capabilities from inception.

ACKNOWLEDGMENTS

This work was supported by the Department of Psychiatry (PRIDE grant), NIH Autism Center of Excellence Award (NICHD P50 HD093074), SFARI (Simons Foundation), ONR, NGA, ARO, and NSF. G.D. is on the Scientic Advisory Boards of Janssen Research and Development, Akili, Inc., LabCorp, Inc. and Roche Pharmaceuticals, has received grant funding from Janssen Research and Development, L.L.C., received royalties from Guildford Press and Oxford University Press, consults for Apple, and is a member of DASIO, L.L.C. G.S. is on the Board of Surgical Information Sciences, L.L.C., consults for Apple, and is a member of DASIO, L.C.C. The work here reported is carried out independently and has not been influenced by these activities/organizations or financial support. All parents/legal guardians of participants gave written, informed consent, and the study's protocol was approved by the Duke University School of Medicine Institutional Review Board.

REFERENCES

[1] *Diagnostic and Statistical Manual of Mental Disorders, 4th Edn*. American Psychiatric Association, 2000.

[2] G. Dawson, A. Meltzoff, J. Osterling, J. Rinaldi, and E. Brown, "Children with autism fail to orient to naturally occurring social stimuli," *Journal of Autism and Developmental Disorders*, no. 28, pp. 479–485, 1998.

[3] G. Dawson, K. Toth, R. Abbott, *et al.*, "Early social attention impairments in autism: Social orienting, joint attention, and attention to distress," *Developmental Psychology*, no. 40(2), pp. 271–283, 2004.

[4] K. Chawarska, S. Macari, and F. Shic, "Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders," *Biological Psychiatry*, no. 74(3), 195-203.

[5] A. Klin, S. Shultz, and W. Jones, "Social visual engagement in infants and toddlers with autism: early developmental transitions and a model of pathogenesis," *Neurosci Biobehav Rev.*, no. 50, pp. 189–203, 2014.

[6] E. Werner, G. Dawson, J. Osterling, *et al.*, "Brief report: Recognition of autism spectrum disorder before one year of age: a retrospective study based on home videotapes," *Journal of Autism and Developmental Disorders*, no. 30(2), pp. 157–162, 2000.

[7] J. Constantino, S. Kennon-McGill, C. Weichselbaum, N. Marrus, A. Haider, A. Glowinski, S. Gillespie, C. Klaiman, A. Klin, and J. W., "Infant viewing of social scenes is under genetic control and is atypical in autism," *Nature*, vol. 547, pp. 340–344, July 2017.

[8] 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," *V. Arch Gen Psychiatry*, no. 59, pp. 809–816, 2002.

[9] M. Murias, S. Major, K. Davlantis, L. Franz, A. Harris, B. Rardin, M. Sabatos-DeVito, and G. Dawson, "Validation of eye-tracking measures of social attention as a potential biomarker for autism clinical trials," *Autism Res.*, no. 11(1), pp. 166–174, 2017.

[10] C. Norbury, J. Brock, L. Cragg, S. Einav, H. Griffiths, and K. Nation, "Eye-movement patterns are associated with communicative competence in autistic spectrum disorders," *Journal of Child Psychology and Psychiatry*, no. 50(7), pp. 834–842, 2009.

[11] K. Pierce, D. Conant, R. Hazin, R. Stoner, and J. Desmond, "Preference for geometric patterns early in life as a risk factor for autism," *Archives of General Psychiatry*, no. 68(1), pp. 101–109, 2011.

[12] F. Shic, J. Bradshaw, A. Klin, B. Scassellati, and K. Chawarska, "Limited activity monitoring in toddlers with autism spectrum disorder," *Brain Research*, no. 1380, pp. 246–254, 2011.

[13] J. Kirchner, A. Hatri, H. Heekeren, and I. Dziobek, "Autistic symptomatology, face processing abilities, and eye fixation patterns," *Journal of Autism and Developmental Disorders*, no. 41(2), pp. 158–167, 2011.

[14] L. Shi, Y. Zhou, J. Ou, J. Gong, S. Wang, and X. Cui, "Different visual preference patterns in response to simple and complex dynamic social stimuli in preschool-aged children with autism spectrum disorders," *PLOS ONE*, no. 10(3), 2015.

[15] J. Swettenham, S. Baron-Cohen, T. Charman, A. Cox, G. Baird, A. Drew, L. Rees, and S. Wheelwright, "The frequency and distribution of spontaneous attention shifts between social and nonsocial stimuli in autistic, typically developing, and nonautistic developmentally delayed infants," *Journal of Child Psychology and Psychiatry*, no. 39, pp. 747–753, 1998.

[16] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: a video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, 2015.

[17] A. N. Esler, V. Bal, W. Guthrie, A. Wetherby, S. Ellis Weismer, and C. Lord, "The Autism Diagnostic Observation Schedule, Toddler Module: Standardized severity scores," *J Autism Dev Disord*, vol. 45, pp. 2704–2720, 2015.

[18] D. Robins, K. Casagrande, M. Barton, *et al.*, "Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f)," *PEDIATRICS*, no. 133(1), pp. 37–45, 2014.

[19] K. Campbell, K. Carpenter, S. Espinosa, *et al.*, "Use of a digital modified checklist for autism in toddlers revised with follow-up to improve quality of screening for autism," *The Journal of Pediatrics*, no. 183, pp. 133–139, 2017.

[20] J. Hashemi, K. Campbell, K. Carpenter, A. Harris, Q. Qiu, M. Tepper, S. Espinosa, J. Schaich-Borg, S. Marsan, R. Calderbank, J. Baker, H. Egger, G. Dawson, and G. Sapiro, "A scalable app for measuring autism risk behaviors in young children: A technical validity and feasibility study," *MobiHealth*, October 2015.

[21] K. Gotham, S. Risi, A. Pickles, *et al.*, "The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity," *Journal of Autism and Developmental Disorders*, no. 37(4), pp. 613–627, 2007.

[22] E. Mullen, *Mullen scales of early learning*. Circle Pines, MN: American Guidance Service Inc, 1995.

[23] "Duke A+ study: A research study for children between ages 1 and 7 years, https://autismcenter.duke.edu/research/duke-study-research-study-children-between-ages-1-and-7-years,"

[24] M. Murias, S. Major, S. Compton, J. Buttinger, J. Sun, J. Kurtzberg, and G. Dawson, "Electrophysiological biomarkers predict clinical improvement in an open-label trial assessing efficacy of autologous umbilical cord blood for treatment of autism," *Stem Cells Translational Medicine*, 2018.

[25] K. Campbell, K. Carpenter, J. Hashemi, S. Espinosa, S. Marsan, J. Schaich-Borg, Z. Chang, W. Qiu, S. Vermeer, M. Tepper, J. Egger, J. Baker, G. Sapiro, and G. Dawson, "Computer vision analysis captures atypical social orienting in toddlers with autism," *Autism: International Journal of Research and Practice*, pp. 1–10, 2018.

[26] E. J. Jones, K. Venema, R. Earl, R. Lowy, K. Barnes, A. Estes, G. Dawson, and S. Webb, "Reduced engagement with social stimuli in 6-month-old infants with later autism spectrum disorder: A longitudinal prospective study of infants at high familial risk," *Journal of Neurodevelopmental Disorders*, vol. 8, 2016.

[27] F. De la Torre, W.-S. Chu, X. Xiong, *et al.*, "Intraface," *11th IEEE International Conference and Workshops on Face and Gesture Recognition*, pp. 1–8, 2015.

[28] J. Hashemi, G. Dawson, K. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. Baker, H. Egger, and H. Sapiro, "Computer vision analysis for quantification of autism risk behaviours," *IEEE Transactions on Affective Computing*, 2018, to appear.

[29] T. Baltruaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 110, 2016.

[30] T. Baltruaitis, A. Zadeh, Y. Chong Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.
[31] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, no. 24(6), pp. 381–395, 1981.
[32] M. Lucas da Silva, D. Goncalves, T. Guerreiro, and H. Silva, "A web-based application to address individual interests of children with autism spectrum disorders," *Procedia Computer Science*, vol. 14, pp. 20–27, 2012.
[33] M. Lucas da Silva, , H. Silva, and T. Guerreiro, "Inducing behavior change in children with autism spectrum disorders by monitoring their attention," *Proceedings of the International Conference on Physiological Computing Systems*, vol. 1, pp. 131–136, 2014.
[34] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3:19, June 2018.
[35] K. Pelphrey, N. Sasson, S. Reznick, G. Paul, B. Goldman, and J. Piven, "Visual scanning of faces in autism," *Journal of Autism and Developmental Disorders*, vol. 32, no. 4, pp. 249–61, 2002.
[36] N. Merin, G. Young, S. Ozonoff, and S. Rogers, "Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at risk for autism from comparison infants," *J Autism Dev Disord*, no. 37, pp. 108–121, 2006.
[37] H. Egger, G. Dawson, J. Hashemi, K. Carpenter, S. Espinosa, K. Campbell, S. Brotkin, J. Shaich-Borg, Q. Qiu, M. Tepper, J. Baker, R. Bloomfield, and G. Sapiro, "Automatic emotion and attention analysis of young children at home: A researchkit autism feasibility study," *npj Nature Digital Medicine*, June 2018.
[38] G. Dawson, K. Campbell, J. Hashemi, S. Lippmann, V. Smith, K. Carpenter, H. Egger, S. Espinosa, S. Vermeer, J. Baker, and G. Sapiro, "Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder," 2018, under review.
[39] Q. Qiu *et al.*, "Low-cost gaze and pulse analysis using realsense," *MobiHealth*, 2015.
[40] Z. Chang, Q. Qiu, and G. Sapiro, "Synthesis-based low-cost gaze analysis," *International Conference on Human-Computer Interaction*, July 2016.

Matthieu Bovery is a student in the Electrical Engineering Department at the Ecole Normale Supérieure (ENS) Paris-Saclay. He has completed his first year of master degree in Electrical Engineering, Robotics and Electronics (M1 E3A) and is currently working as an intern researcher in electrical engineering and computer science for Duke University. His research topic includes machine learning, computer vision and applied computer science.

Geraldine Dawson is Professor in the Departments of Psychiatry and Behavioral Sciences, Pediatrics, and Psychology & Neuroscience at Duke University. She is Director of the Duke Center for Autism and Brain Development, an interdisciplinary autism research and treatment center. Dawson is Director of an NIH Autism Center of Excellence Award at Duke focused on understanding early detection, neural bases, and treatment of autism and ADHD. Dawson has published extensively on early detection, brain function, and treatment of autism.

Jordan Hashemi received his BEng from the Department of Biomedical Engineering and his MSc from the Department of Electrical Computer Engineering at the University of Minnesota in 2011 and 2013, respectively. He is currently working towards the PhD degree in electrical and computer engineering at Duke University. He was awarded the Kristina M. Johnson Fellowship award in 2015. His current research interests include applied computer vision, machine learning, and behavioral coding analysis. He is a student member of IEEE.

Guillermo Sapiro was born in Montevideo, Uruguay, on April 3, 1966. He received his B.Sc. (summa cum laude), M.Sc., and Ph.D. from the Department of Electrical Engineering at the Technion, Israel Institute of Technology, in 1989, 1991, and 1993 respectively. After post-doctoral research at MIT, Dr. Sapiro became Member of Technical Staff at the research facilities of HP Labs in Palo Alto, California. He was with the Department of Electrical and Computer Engineering at the University of Minnesota, where he held the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. Currently he is a James B. Duke Professor with Duke University. He works on theory and applications in computer vision, computer graphics, medical imaging, image analysis, and machine learning. He has authored and co-authored over 400 papers in these areas and has written a book published by Cambridge University Press, January 2001. He was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Post-Doctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the National Science Foundation Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010. He received the test of time award at ICCV 2011, was the founding Editor-in-Chief of the SIAM Journal on Imaging Sciences, and is a Fellow of IEEE, SIAM, and the American Academy of Arts and Sciences.

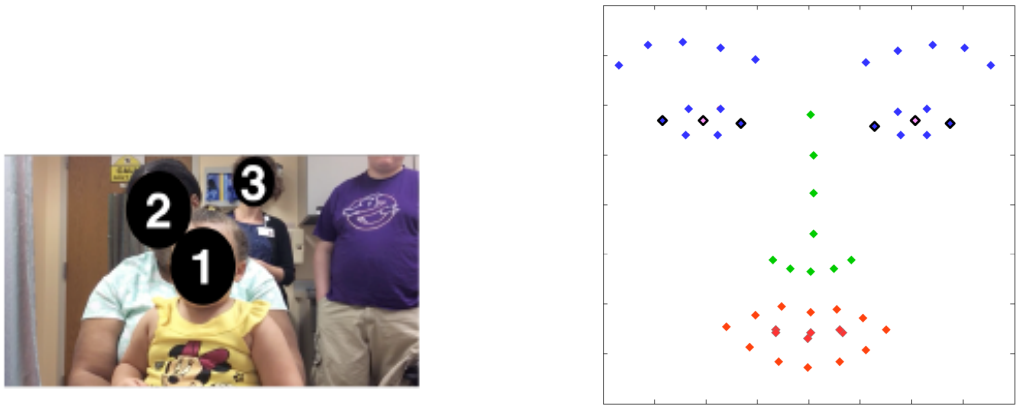

Fig. 1. Screenshot of the recorded video from the front facing tablet's camera (left), and example of automatic facial landmarks used for attention detection (right). The child (1) is sitting on the caregiver's (2) lap, while the practitioner (3) is standing behind in this example. The six outlined automatically detected landmarks (in black) are the ones used for measuring the direction of the attention.

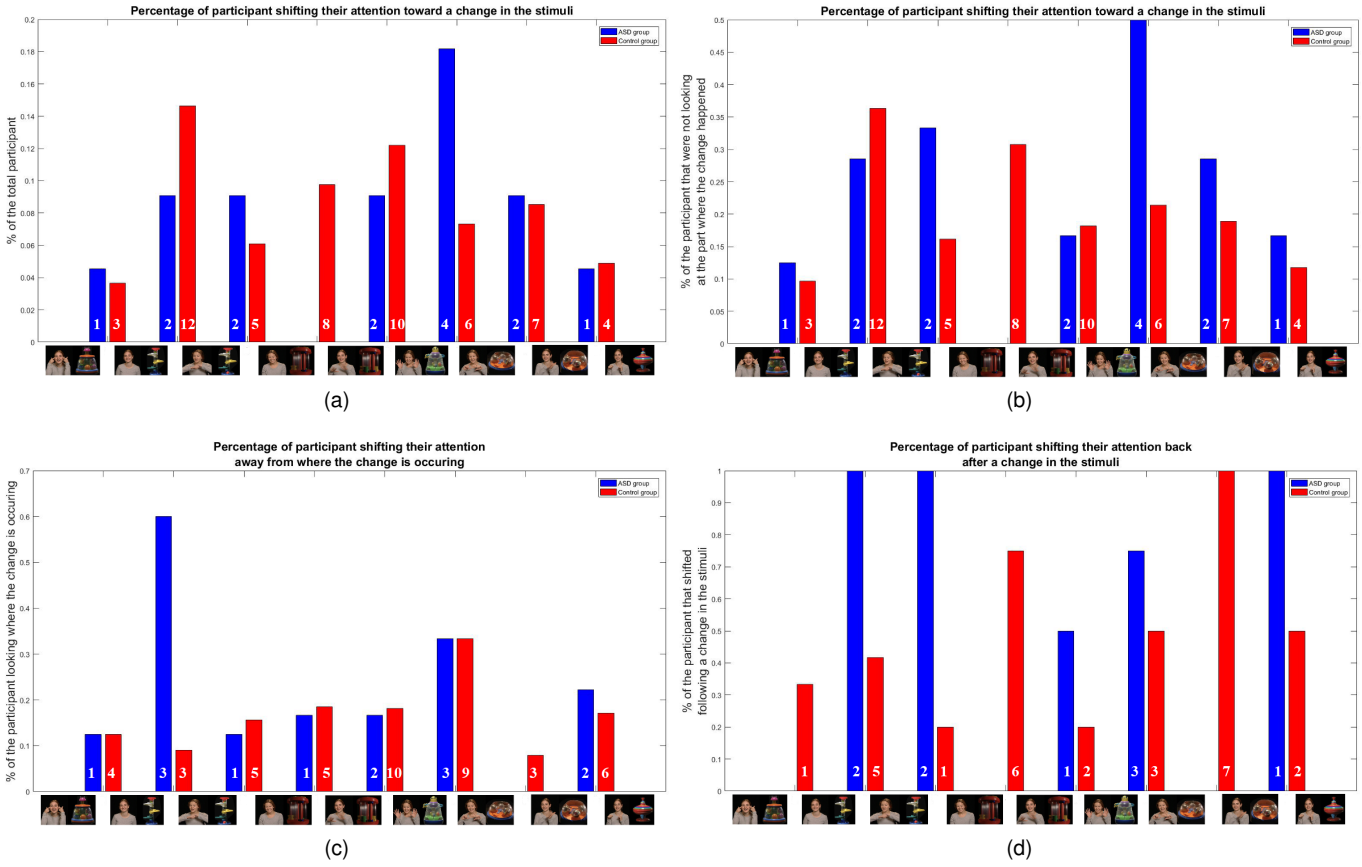

Fig. 8. Illustration of use of the proposed computational approach for monitoring the dynamic change in attention, as a response to changing stimulus. The figures show percent of participants performing certain dynamic pattern of shift of attention between the social and the non-social halves of the stimulus; see caption above each figure and text for the particular pattern. The total number of subjects per class/stimulus switch is indicated by the numbers in each bar. These results need to be further studied in large populations for their statistical power. See text for details and Figure 2 for larger stimulus frames.