HHS Public Access

Author manuscript

IEEE Trans Affect Comput. Author manuscript; available in PMC 2022 April 08.

Published in final edited form as:

IEEE Trans Affect Comput. 2021 ; 12(1): 215–226. doi:10.1109/taffc.2018.2868196.

Computer Vision Analysis for Quantification of Autism Risk Behaviors

Jordan Hashemi [Student Member IEEE],
Department of Electrical and Computer Engineering, Duke University, Durham, NC

Geraldine Dawson,
Department of Psychiatry and Behavioral Sciences, Duke Center for Autism and Brain Development, and the Duke Institute for Brain Sciences, Durham, NC

Kimberly L. H. Carpenter,
Department of Psychiatry and Behavioral Sciences, Duke Center for Autism and Brain Development, and the Duke Institute for Brain Sciences, Durham, NC

Kathleen Campbell,
Department of Psychiatry and Behavioral Sciences, Durham, NC

Qiang Qiu,
Department of Electrical and Computer Engineering, Duke University, Durham, NC

Steven Espinosa,
Department of Electrical and Computer Engineering, Duke University, Durham, NC

Samuel Marsan,
Department of Psychiatry and Behavioral Sciences, Durham, NC

Jeffrey P. Baker,
Department of Pediatrics, Duke Health, Durham, NC

Helen L. Egger,
Department of Child and Adolescent Psychiatry, NYU Langone Health, New York, NY. She performed this work while at Duke University

Guillermo Sapiro [Fellow IEEE]
Department of Electrical and Computer Engineering, Duke University, Durham, NC

Abstract

Observational behavior analysis plays a key role for the discovery and evaluation of risk markers for many neurodevelopmental disorders. Research on autism spectrum disorder (ASD) suggests that behavioral risk markers can be observed at 12 months of age or earlier, with diagnosis possible at 18 months. To date, these studies and evaluations involving observational analysis tend to rely heavily on clinical practitioners and specialists who have undergone intensive training to be able to reliably administer carefully designed behavioural-eliciting tasks, code the resulting behaviors, and interpret such behaviors. These methods are therefore extremely expensive, time-

Corresponding author jordan.hashemi@duke.edu.

intensive, and are not easily scalable for large population or longitudinal observational analysis. We developed a self-contained, closed-loop, mobile application with movie stimuli designed to engage the child's attention and elicit specific behavioral and social responses, which are recorded with a mobile device camera and then analyzed via computer vision algorithms. Here, in addition to presenting this paradigm, we validate the system to measure engagement, name-call responses, and emotional responses of toddlers with and without ASD who were presented with the application. Additionally, we show examples of how the proposed framework can further risk marker research with fine-grained quantification of behaviors. The results suggest these objective and automatic methods can be considered to aid behavioral analysis, and can be suited for objective automatic analysis for future studies.

Keywords

Computer vision; autism; behavior elicitation; behavior coding; mobile-health

1 INTRODUCTION

OBSERVATIONAL behavior analysis has played, and still plays, a key role for gaining insight into mechanisms and risk markers of impairing neurodevelopmental disorders such as autism spectrum disorder (ASD). The gold standard observational tool for ASD diagnosis, the Autism Diagnostic Observational Schedule (ADOS, [1]), requires several observational coding components, as does an assessment of early risk markers of ASD, the Autism Observational Scale for Infants (AOSI, [2]). Retrospective behavioral analysis of home videos helped discover early risk markers involving diminished social engagement and joint attention in children who were later diagnosed with ASD [3], [4], [5], [6], [7]. Research studies have documented several other early behavioral risk markers of ASD that emerge within the first months of life; these include atypical visual attention related tasks, such as difficulty responding or orienting to a name-call when engaged with an activity, difficulty disengaging from a stimulus when a competing one is presented, and non-smooth visual tracking [8], [9], [10]. Additionally, children with ASD may also exhibit atypical social behaviors such as, decreased expression of positive affect, decreased frequency of social/shared smiles, decreased frequency of gaze to faces, and decreased eye contact [10], [11], [12]. These advancements in understanding early behavioral development aid in the development of tools for screening, diagnosing, and monitoring ASD.

Studies and evaluations involving behavioral analysis tend to rely heavily on medical practitioners and specialists who have undergone intensive training to be able to reliably administer the eliciting tasks and then code and interpret the observed behaviors. In behavioral studies, practitioners tend to code behaviors based on clinical judgment and thus tend to code behaviors more subjectively and at a lower granularity than is possible in computer analysis. In retrospective analysis, specialists can potentially go frame-by-frame to hand code behaviors; this is not only burdensome, but is not easily scalable for big data studies aimed at discovering or refining behavioral risk markers or for longitudinal tracking. As technology has advanced, new tools have emerged to assist in automatic and semi-automatic behavioral coding in infants and toddlers. Eye-tracking is a great example,

where technological advancements have made major impacts in the understanding of ASD behavioral development. Researchers are able to automatically code behavior from gaze and analyze it at fine-grained scales, leading to novel understandings of development of ASD, such as decreased preferential attention to eye and mouth regions of a face and impaired oculomotor control [13], [14]. However, standard eye-tracking systems for toddlers are still very constrained, specialized, and expensive, making the availability and reach of these systems limited. Research using automatic behavioral coding in less constrained settings, from schools to homes, have also been promising, where for example researchers explore tools and a dataset to develop and evaluate social and communicative behaviors relevant to child-adult interaction sessions [15]. In another recent work, researchers automatically encoded motor movements during mother-infant interactions to explore quality of interaction [16]. There has also been some important progress in augmenting the coding for visual attention tasks in infants with the use of just a single consumer grade camera [17], automating coding of head movement dynamics while watching movies of nonsocial and social stimuli [18], and monitoring facial expression imitation of a child as he/she interacts with a robot [19].

In this work we concentrate on the use of ubiquitous devices, like smart phones and tablets, developing a self-contained solution that does not need any additional hardware at all. Focusing on unconstrained and low-cost setup requirements is especially important since many middle and low resource communities lack access to specialists in ASD, thus lack access to any sort of evaluation. Additionally, unconstrained setups allow for behavioral monitoring in more naturalistic environments, such as at home. There is a desire for universal and well-validated behavioral tools to further research on early risk marker detection in ASD. The tools have to be universal in the sense that they are consistent and accessible across different user groups; and have to be validated in the sense that they have to agree with trained specialists, potentially helping to discover new biomarkers as further discussed in this paper. It is worth noting that ASD diagnosis involves much more than the detection of risk markers, but furthering the accessibility and development of screening tools for identifying toddlers that might be at risk and informing caregivers, is of great value. Although risk behaviors can be observed as early as at 12 months of age leading to diagnosis at 18 months, the average age of diagnosis in [sic] the United States is around ~4−5 years old [20]. Not only can intensive early intervention provide long-term improvements for the child, but also starting intervention before the full syndrome is present can have an even greater effect on outcomes [21], [22], [23].

Towards these challenges and opportunities, our interdisciplinary team developed a self-contained mobile application with movie stimuli designed to elicit and quantify specific ASD-related behavioral responses in toddlers. In a long-term effort, this project aims to develop low-cost, automatic, and quantitative tools that can be used by researchers and general practitioners in general settings (such as clinics or schools, and potentially by caregivers at home) to identify toddlers at risk for ASD or other developmental disorders.

We have developed a novel application of displaying movie stimuli on a mobile device which were expertly designed to capture the toddler’s attention and elicit behaviors relevant to early risk markers of ASD, including orienting to name call, social referencing,

smiling while watching the movie stimuli, pointing, and social smiling (smiling while social referencing) [10], [11], [24], [25]. Using the front facing camera of the mobile device, we capture the toddler and automatically code these event-related behaviors with computer vision analysis tools. The current study presented in this paper validates our automatic codings of engagement, name-call responses, and emotion with hand codings from specialists on a diverse population including toddlers with ASD and without ASD (non-ASD). We also present examples of behavioral analysis extracted from the validated automatic methods. In a parallel publication [26], we study the feasibility of the tools and paradigm here introduced and described in the home environment.

It is important to stress that while the framework here introduced can be used in any environment, from clinics to homes, it is not a passive monitoring system but a user friendly and active one. In passive systems, e.g., those used to monitor heart conditions via wrist watches, the signal to noise ratio is very poor, in particular when aiming at capturing highly accurate and risk informative behaviors within a short time span. With a closed-loop and self-contained system like the one here proposed, where carefully designed short and entertaining movie stimuli elicit behaviors (as in ADOS or AOSI standard approaches, but without the human in the loop), we obtain a much more valuable and interpretable signal.

2 DATA COLLECTION

2.1 Setup and participants

The work here presented is an extension of the preliminary work reported in a conference proceedings [27]. In addition to providing more details missing in the limited-space conference paper, we extend the cited work in a multitude of ways, including developing an automatic method for name-call prompt detection (critical due to the demonstrated relevance of this important stimulus), developing methods for engagement and attention analysis, increasing and diversifying the sample populations, including an additional movie stimuli for analysis, validating the proposed methods at a per-video and at a frame-by-frame basis, and demonstrating how the developed methods can be used to analyze additional behaviors relevant to risk markers such as head motion. The study was carried out in a pediatric care clinic with the approval of the Duke Health Institutional Review Board. Caregivers and toddlers visiting the clinic for an 18 or 24 month well-child visit, where all toddlers in the clinic are screened for ASD with The Modified Checklist for Autism in Toddlers – Revised with Follow-up (M-CHAT-R/F) [28], were invited to participate in the study. Children with known hearing or vision impairments and caregivers who could not complete consenting in English were excluded. During the well-child visit, caregivers were asked to hold the toddler on his/her lap while an iPad (4^{th} *generation*) was placed on a stand at the toddler's eye-level and set about 1 meter away, see Figure 1. To minimize distractions, the study was carried out in the doctor's office while the caregivers and toddler were waiting for the doctor, and all other family members, persons, and the practitioner were asked to stand behind the caregivers and toddler. Reducing distractions in in general important for behavioral tests. Our companion at-home study, [26], shows that the algorithms here presented can be applied to studies carried-out in more natural environments as well. Caregivers were told they could interact with their toddler for the first 45 seconds of the session while a

'mirror' was presented on the screen (child can see him/herself on the screen, which is showing the tablet's camera input). After this time, the caregivers were asked to remain quiet and not direct their child's behavior or attention once the movie stimuli began. For the rest of the session, the iPad displayed different carefully designed movie stimuli on the screen (one stimulus at a time, with nothing else shown on the screen), while simultaneously recording the child's face via the front facing camera at 1280×720 resolution and 30 frames per second. If a child screened positive on the M-CHAT-R/F, or a caregiver or clinician expressed concerns about ASD during the visit, then the child received gold standard diagnostic testing with the Autism Diagnostic Observational Schedule – Toddler Module with a child psychologist for final diagnosis [29]. As a validation set for the developed computational methods, we consider a total of 33 participants (18 non-ASD and 15 ASD) ages 16 – 31 months enrolled in our study. The selection of participants for the study was based on age distribution to represent the range of ages for both non-ASD and ASD groups. Full demographics are shown in Table 1.

2.2 Movie stimuli

A series of stimuli were presented on the iPad screen, consisting of short developmentally appropriate movies designed to engage the child's attention and elicit specific behavioral and social responses (such as smiling at specific events during the movie, social referencing, and social or shared smiling). The movies incorporated a 'name-call' prompt where during the movie the practitioner, who was standing behind the child and caregiver, would call the child's name loudly while the movie is still being displayed on the screen (this is of course easy to incorporate in an automatic fashion, e.g., via bluetooth speakers). Screenshots of the stimuli and the recording from the iPad are shown in Figure 1. We considered the set of 3 movie stimuli, namely:

Bubbles.—Bubbles are presented at random and moving throughout the frame. A 'name-call' is presented once during the movie. The total duration is 30 seconds.

Bunny.—A mechanical toy bunny is presented on one side of the screen and hops horizontally towards the other side, which contains a group of toy vegetables. As the bunny reaches the midpoint of its path, an animal puppet is introduced and temporarily disrupts the bunny's path. A 'name-call' is presented once during the movie. The total duration is 66 seconds.

Puppets.—Two animal puppets interact while building a block tower together and then knock it down. The tower is built three times and knocked down twice. A 'name-call' is presented once during the movie. The total duration is 68 seconds.

3 METHODS

We developed automatic video and audio analysis methods to study behaviors and child responses related to attention and facial expression. During the movie stimuli, the text 'Name' appeared in the upper left corner to prompt the practitioner to call the child's name once loudly (see previous section). Although it is known when the text appeared on the screen in relation to the movie, there is a need for automatic detection of the exact time the

practitioner said the child's name. To this end, we also developed an automatic name-call detection based on the recorded audio. To validate the automatic methods, expert human raters coded emotion, name-call prompts, and name-call responses. While the computer vision and machine learning communities have now very advanced ground-truth data, including for human emotions, this is restricted to mostly healthy adults. It is therefore imperative to validate results with the population of interest, since it is well documented that children, and also children with developmental disorders, exhibit very different facial expressions when compared with such standard database populations.

3.1 Automatic coding

3.1.1 Name-call prompt detection—Since we know when in the movie stimuli the practitioner is prompted to say the child's name for a name-call, a 4 second window of the recorded audio data is extracted around each of the known name-call text prompts starting 1 second before the time when the text prompt occurred. Focusing on the power spectrum density (psd) of the extracted audio signal in the 300 – 1,200 Hz frequency range, we estimate when the adult practitioner said the child's name by finding the time corresponding to maximum speech energy, see Figure 2.

3.1.2 Face pre-processing—A facial landmark detector and tracker was deployed to track 49 facial landmark points [30]. While other algorithms could be used without affecting the proposed paradigm, this one was found to be very efficient for the desired tasks. Using a subset of the landmark points, namely the inner and outer eye points and the nose spline (see Figure 1), faces were aligned and normalized to a frontal canonical face model by finding the affine transformation between them. In turn, this transformation also represents the head pose estimation. The facial landmark detector requires both eyes to be present in the frame, thus the range of estimated yaw pose θ_{yaw} is {−45°, +45°} (left-right head orientation). We assume that the toddler's head orientation is directly correlated to if he/she is watching towards the stimuli. This assumption is supported by the 'center bias' property that is well established in gaze estimation literature [31], [32]. Engagement when the toddler is watching towards the movie stimuli is defined by frames when the toddler exhibits yaw poses with magnitudes less than 20°.

3.1.3 Head movement and turn detection—We estimate the child's head movement by tracking the distances and pixel-wise displacements of central facial landmarks. We record the frame-by-frame displacements of landmarks around the nose, namely the two outer eye landmarks and the lowest nose landmark shown in Figure 1. The magnitudes of these displacements are heavily dependent on the distance the child is away from the camera. Thus these displacements need to be normalized with respect to the child's distance from the camera. If depth information were available, this would be a trivial task; however, since it is not, we normalize the displacements with respect to the distance between the child's eyes, keeping in line with the use of only available and ubiquitous hardware. At any given time point, the displacements from the nose landmark are normalized by a ±1 second windowed-average Euclidean distance between the eyes.

Since the practitioner and caregiver are located behind the child, the child must transition his/her face from looking at the screen to looking behind him/her in order to perform a head turn (in response to name calling or social referencing for example). To detect head turns and distinguish between a head turn and just an occlusion of the face, we tracked yaw pose changes and defined two rules: to initiate a head turn the pose had to go from a frontal to one extreme head pose position (left or right); to complete a head turn the pose then had to come back from the same extreme position to a frontal position. More formally, to initiate a head turn the yaw pose had to change from a frontal position $\theta_{yaw} \in \{-20°, +20°\}$ to an extreme $|\theta_{yaw}| > 35°$ within a half-second window. Then to complete a head turn, the yaw pose had to change from the same extreme position back to a frontal position, $|\theta_{yaw}| \leq 20°$, within a half-second window. These time intervals represent a child performing a quick head turn (for example responding to a name-call prompt or social referencing). An example of this procedure is shown in Figure 3.

3.1.4 Pose-invariant emotion classification—

To analyze children's behavior in such an unconstrained setting, there is a need for an emotion classification method that can handle faces across varying poses. We employ a modified version of the robust pose-invariant method described in [33], where we first learn a cross-modality and pose-invariant dictionary. This learned dictionary creates a mapping between facial information from both 2D and 3D modalities and is then able to infer discriminative facial information even when only 2D facial information is available at testing time (see [33] for details). For training we use data from Binghamton University 3D Facial Expression database [34], synthesize face images with varying poses, and extract local binary patterns (LBP, [35]) and the distances between a subset of facial landmarks as features to learn the cross-modality and pose-invariant dictionary. Using the inferred discriminative 3D and frontal 2D facial features, we train a multi-class support vector machine [36] to classify four different facial expressions (angry, sad, happy, and neutral) provided by the standard Cohn-Kanade database (CK+, [37]). This method focuses on pose-invariant facial expression recognition in images, and outperformed previous state-of-the-art methods.

3.2 Human coding

Expert human raters coded facial expressions of emotion, head turns, and instances when a name-call prompt occurred. To code emotion, expert human raters were trained to detect facial action units (AUs) from the baby Facial Action Coding system (Baby - FACs, [38]). Since the movie stimuli were designed to elicit positive emotion and smiling responses, the expert human raters focused on coding 'Happy' emotion and coded it when activation of the zygomaticus major (AU12) occurred and coded 'Other' otherwise. Expert human raters coded 'Not Visible' when the childs face was covered, out of the field of view, or when more than half of the face was not visible due to head turning away from the camera. Head turning was also coded when a child turned his/her head to look at the practitioner or caregiver (since they were located behind the child). Raters were not blind to diagnostic group, but were blind to stimuli and videos were muted during coding to prevent the influence of vocalizations on the coding of emotions.

Additionally, all name-call prompts were coded separately by a practitioner who marked the frames in the video when the child's name was called. Coding by the expert human raters and practitioner were performed using the Nodlus Observe XT software Version 11.0 [39].

4 VALIDATION RESULTS AND ANALYSIS

We now present validation results on the automatic coding of engagement, name-calls including detecting the timing of the prompt and the children's responses, and children's emotions. We also present examples of uses of the validated methods for general behavioral analysis. For this work we use two-way, random, consistency intra-class correlation coefficient (ICC) scores and 95% confidence intervals (CI) to report inter-rater reliability performance; where ICC scores were classified as weak (ICC < 0.4), moderate (0.4 ≤ ICC≤ 0.75), and excellent (ICC > 0.75) [40], [41]. Precision, recall, and F1 scores are also employed for validation of emotion classification. While precision considers the accuracy of the decisions made by the automated methods, recall considers the fraction of correct decisions made by the automated methods out of all the decisions made by the expert human raters, and F1 is the harmonic mean of precision and recall [42]. The F1 score conveys the balance between precision and recall scores, and is useful when dealing with uneven distributions and to report a single performance measure. Statistical analysis for agreements were conducted in R Version 3.4.3 [43], using the *irr* package [44] to compute ICC scores and the *ROCR* package [45] for precision, recall, and F1 computations.

Across all participants, 99 video recordings were considered (Table 2). Two expert human raters were first trained on a reliability dataset (separate from analyzed participants) until they reached excellent agreement, ICC > 0.75. Then, a single expert human rater coded all 99 video recordings, while the second rater coded ~20% of them to verify ongoing inter-rater reliability. Overall agreement between the human raters for coding of engagement, facial expression, and social referencing achieved an ICC score of 0.84 with 95% CI of 0.76–0.95. For the validation here reported, we compare codings between the automated methods and the expert human rater who coded all of the video recordings.

4.1 Validation results

4.1.1 Engagement detection

Inter-rater reliability performance of detecting the total time the child was engaged was assessed on a per-video basis, Table 3. Reliability between the automatic methods and the expert human rater was excellent, achieving an ICC score of 0.85; while the reliability for the sub groups of participants with and without ASD was also excellent, achieving ICC scores of 0.81 and 0.89 respectively. Since our automatic methods code per-frame, we also analyzed accuracy on a per-frame basis. Across the 161,296 coded frames for engagement (~89 minutes), our methods matched with the expert human rater on 90% of them (with most of the differences located at the start or end of a new behavior). Accuracy of coded frames from the sub groups of participants with and without ASD exhibited similar results. Across 74,670 coded frames from videos of participants with ASD, the agreement accuracy was 85%; where across the 86,626 coded frames from videos of participants without ASD, agreement accuracy was 94%.

4.1.2 Name-call prompt and response detection—

To fully assess a child's behavior to a name-call, we first detect the time a name-call happened and then detect whether or not the child turned his/her head to orient to the name-call. The practitioner manually marked every name-call prompt across all participants by marking the frame in the video when the child's name was called. For all 99 name-call prompts that were marked, the automatic name-call prompt detection was able to detect every name-call by the practitioner. Fitting an exponential model to the data, we saw a mean absolute time difference of just 0.21 seconds (Figure 4a).

Child responses after name-call prompts were also coded, where a head turn response was coded if the child performed a head turn within 5 seconds of the name being called. Compared to the expert human rater, the automatic head turn detection method correctly identified 84% of the child responses (46 of the 60 head turn responses, see Section 5.1 for a discussion on this, and 37 of the 39 non head turn responses). Overall it achieved an excellent reliability with an ICC of 0.84 (Table 3), while the reliability of the sub groups of participants with and without ASD was also excellent, achieving 0.80 and 0.86 ICC scores respectively. By fitting an exponential model to the 46 correctly detected head turns, the mean absolute time difference between the automatic method and the expert human rater was 0.22 seconds (Figure 4b).

4.1.3 Emotion classification—

The automatic emotion classification method outputs ℓ_1 normalized probability weights across emotions. To compare it to the expert human rater, we grouped the emotions into the categories coded by the rater. Namely, 'Happy' considers automatic coding of happy (the key emotion the designed movies is expected to elicit) while 'Other' considers angry, sad, and neutral. Automatic coding of emotion was done at a much finer scale than that of the expert human rater, where the automatic coding estimated probability scores for each emotion and treated each frame independently, whereas the human rater assigned each frame only one emotion and marked time points when the given emotion started and ended. To accommodate these differences, half-second filtering of probability scores and max-voting in ± half-second windows were performed on the automatic coding to determine which emotion was dominantly expressed when the face was detected.

Inter-rater reliability performance of quantifying the total time the child was exhibiting Happy was assessed on a per-video basis, where the inter-rater reliability between the automatic methods and the expert human rater was excellent, achieving an ICC of 0.90 (Table 3). The reliability for the sub groups of participants with and without ASD was also excellent, achieving ICC scores of 0.90 and 0.89, respectively. Performance of the automatic methods was also validated on a per-frame basis. 136,450 frames (~75 minutes) were coded for emotion across all the participants; Table 4 shows the precision, recall, and F1-scores of the automatic emotion classification method using the expert human rater as ground truth. Overall, the automatic method achieved high precision, recall, and F1 scores: 0.89, 0.90, and 0.89 respectively.

4.2 Behavioral analysis

4.2.1 Extracted behaviors—We now present multiple examples of ASD-risk related behaviors that can be automatically extracted using the above presented and validated methods, including responses from name-call, engagement, and emotion. During name-call responses it is important to not only code if the child oriented to a name-call but also the latency between when the child oriented and when the child's name was called. Examples of head turn responses to 3 name-call prompts for a non-ASD and ASD participant are shown in Figure 5. The non-ASD participant not only oriented to all 3 name-call prompts, but all head turns exhibited less than half a second latency. The ASD participant on the other hand, only oriented to 1 of the name-call prompts (the second one in this case) and exhibited a large latency of greater than one and a half seconds. The full population study of this behavior is reported in [46], see below.

With the validated methods, more in-depth head movement dynamics can be extracted beyond head turning. Figure 6 shows examples of extracted head movement tracks and cumulative head movement (as a simple quantification tool) of three participants while watching the Puppets movie stimuli. Each participant demonstrates a distinct movement profile, varying from sitting still (part. 01) to moving a lot throughout the stimuli (part. 03). This is all automatically measured without any additional hardware, such as motion capture devices.

Emotion related responses to scenes in the movie stimuli, designed to elicit such emotions, as well as spontaneous emotions, are also of great importance for ASD related risk behaviors. Although for the validation of emotion coding each frame was assigned a categorical label, the presented methods provide fine-grained, continuous probability scores for each emotion. Figure 7 shows examples of participants without and with ASD expressing Happy when the bunny hops in the Bunny movie stimuli. Both participants react by expressing Happy during this segment, but the participant with ASD exhibits less instances of (high probability) Happy. In [26] we report further use of the automatic emotion coding here introduced to analyze over 4,000 movies recorded at home, see next.

4.2.2 Extensions—A direct extension of this work is presented in the recent companion clinical publication [46], where these automated methods were used on a study considering 104 toddlers (82 non-ASD and 42 ASD, ages 16–31 months). The work reported that toddlers with ASD not only tend to orient fewer times as a response to the name-call, but the mean latency to orient was significantly longer compared to non-ASD (2.02 vs. 1.06 seconds). Only automatic frame-rate methods, as the one here presented, can detect such important differences and potential risk markers in an easily scalable manner. The engagement coding allows for further quantification of compliance as well as attention span. Using these automatic methods here described and detailed, [46] also reported that there was a significant interaction between diagnostic group and age, with the ASD group showing significantly lower amount of time engaged in the task than the comparison group at older ages only.

Another extension of this work and the exploitation of the algorithms here introduced is presented in our at-home companion study [26], where 1,756 families participated uploading

a total of 4,441. Applying the automatic algorithms described in this paper, 87.6% of the frames were usable, providing rich data at a previously unseen scale for the applications here considered. It was reported, for example, that from the 530 children (ages 12–72 months) who watched the same movie stimuli, children with high risk for ASD (self-reported or due to high M-CHAT-R/F score) exhibited lower mean percentage (± standard error) of positive emotions compared to children not at risk (29.9% ± 1.9 vs. 35.1% ± 1.3, $p<.02$). Additionally it was reported that with increased age, children showed a greater percentage of exhibiting positive emotion while watching the movie stimuli (Pearson Correlation Coefficients with age: total positive emotion 0.208, $p<.0001$). See [26] for numerous additional findings resulting from the application of the algorithms here described to this large data.

These companion clinical papers, using a subset of the same stimuli and algorithms here described, show the scalability of the approach and its usability in diverse environments.

5 DISCUSSION

We have been developing an integrated paradigm where short and entertaining movies, carefully designed to elicit risk behaviors related to ASD or other developmental disorders, are presented on a mobile platform while the device's camera is recording the participant's responses [26], [27], [46]. This work concentrated on the validation of the computer vision components of this paradigm, in particular for ASD risk markers. Contrary to the standards in the computer vision and machine learning related literature, the ground truth used for validation comes from the population under study and the labeling is done by domain experts. The results indicate high agreement between the proposed automatic methods and the expert human coders, and that the automatic methods can be considered to augment behavioral analysis of early risk markers. Furthermore, these validated methods allow for automatic and objective measurements of high granularity and have many potential benefits for future research of risk markers. This scalability and low cost of the paradigm, basically software only due to the ubiquitous presence of mobile devices, opens the door to deployment on longitudinal studies to track behavioral progress and development. The high granularity of the proposed methods can also lead way to refined definitions of risk markers and behavioral trajectories. With these now validated automatic computer vision methods, we presented multiple examples of behaviors that can be extracted. For name-call responses, decisions based on the participant's head turning and the latency to turning after the name has been prompted can be extracted (Figure 5). Attention characteristics and head movement dynamics of the participant throughout the movie stimuli can also be accurately quantified (Figure 6). Emotional responses, elicited or spontaneous, during specific events can also be automatically captured (Figure 7). While these behaviors are known to be relevant for early risk markers in ASD during actual physical interactions with a trained examiner, it remains to be fully verified that they are still present in this setting of watching movie stimuli. With this said, there are indications that ASD/non-ASD differences in response to name-call, attention, and emotion can be elicited and captured from this setting and by automatic methods [18], [26], [46].

While in the work here presented, and as it is very common in the literature, we went from automatic landmark detection to automatic behavioral coding of emotions (and head positions), it is of interest to go directly from landmarks to screening/diagnosis via the training of modern classification methods (see also Figure 6, motion is derived directly from landmarks). This possibility increases with the availability of data, [26], and with the application of the scalable tools here described. This will also help to mitigate cultural and data biases in machine learning methods used to train emotion coding algorithms, and will help in the discover of new biomarkers directly connected to the dynamics of our face and not interpreted emotions. The work recently reported in [47] is also a step supporting this important direction of research.

5.1 Limitations

There are still many challenges when automatically coding behaviors in unconstrained settings, and even though the validation results were strong, we see this as a starting point for the proposed methods. Some of the missed head turn responses during a name-call were due to occlusions of the child's faces right before or after a head turn happened (e.g., due to the child's hands occluding his/her face), or from poor image quality when the child's head is moving too quickly (using high frame rates available on mobile phones will address this). Since we want a simple, easy to administer setup requiring only the integrated camera on the mobile device, and also want to keep the most naturalistic setting possible by not constraining the child, chances of occlusions of the face may persist. Future algorithms should incorporate hand detection, e.g., [48], to assist in handling these cases.

Future algorithms should also incorporate gaze analysis and eye-tracking. As mentioned before, eye-tracking captures fine-grained behaviors and can be used to quantify where in the movie scenes the toddlers are looking and fixation patterns, and can be coupled with head pose to assist in defining when the child is engaged with the movie stimuli. A number of low-cost eye-tracking methods are available [49], [50], [51]. Methods based on RGB-D devices, which output both RGB and depth data, have shown to capture gaze in unconstrained settings and with minimal calibration [52], [53]. These advancements are especially promising since depth sensors are more readily being integrated within mobile devices, including smart phones as in our study [26]. Additionally, recent advances in eye-tracking studies have presented generalizable and scalable methods for screening of neurodevelopmental disorders [54], [55], [56], [57]. While these studies do not focus on toddlers and early screening, they add to the argument in favor of incorporating eye-tracking technology in the presented ASD paradigm.

There were differences for precision and recall scores for facial expression (emotions) classification between the ASD and non-ASD groups, opening room for improvement of the current methods. High precision scores show that the automatic methods accurately made decisions; whereas, the lower recall scores indicate that the methods missed some codings from the expert human rater. One possible explanation for the lower recall scores could come from the training data, since the current methods were trained on images of posed expressions, and as such they perform best at the peak of expressions from participants, hence the high precision scores, and may miss the onset and end of expression.

This is a less critical problem if the biomarkers relate to the presence or absence of the emotion event for certain period of time, without the need to use the exact (frame accuracy) time length. We should note that human experts also disagree on the start and end of facial expressions. Additionally, the current facial expression methods were trained on thousands of images from adults and based on adult facial codings. With the annotated data from this work, advancements in techniques to automatically code facial expression, and recent understandings of facial dynamics of individuals with ASD [47], new emotion classification methods based on spontaneous emotion recognition of toddlers are currently being considered by our team. This can be added to the current trained system via modern domain adaptation machine learning techniques.

6 CONCLUSION

We proposed and validated computer vision methods to automatically code behaviors related to early risk markers of ASD. The algorithms are applied to video recordings from the front camera of a mobile device while the child watched movie stimuli designed to elicit such behaviors. In particular, we focused on automatic methods for quantifying engagement, name-call responses, and emotion responses. We validated the automatic methods using manual coding from an expert human rater on a diverse population of toddlers with and without ASD. Additionally, we showed examples of how the proposed methods can further risk marker research with fine-grained quantification of behaviors. The results suggest these low-cost, objective, and automatic methods can be considered to aid behavioral analysis, and can be suited for objective automatic analysis of large and longitudinal studies.

ACKNOWLEDGMENTS

While many people helped with the research here reported, our biggest thanks to the participants. With their caregivers consent, they contributed at already a very early age to the advance of knowledge and helped to improve the way we address ASD challenges.

The work was supported by Duke University, the Duke Endowment, the Coulter Foundation, NSF, Department of Defense, the Office of the Assistant Secretary of Defense for Research and Engineering, and NIH. Grant funding includes NSF-CCF-13-18168, NGA HM0177-13-1-0007 and HM04761610001, NICHD 1P50HD093074-01, ONR N000141210839, and ARO W911NF-16-1-0088.

Biography

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.

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. She served as the lead autism expert on the development of the first autism screening app to use computer vision analysis and the Autism&Beyond autism screening app.

Kimberly L.H. Carpenter is an Assistant Professor in the Center for Autism and Brain Development at Duke University. She is a translational and clinical neuroscientist studying brain-behavior relationships in the first 5 years of life in children with autism and associated psychiatric comorbidities, such as anxiety and ADHD. The goal of her research is twofold: 1) to understand how alterations in neural systems contribute to the etiology and pathophysiology of psychiatric comorbidities in children with autism, and 2) to develop new technologies that use brain- and behavior-based biomarkers for the early identification of these disorders. Through this work, Dr. Carpenter aims to increase access to, and provide a solid neurobiological foundation for, evidence-based screening, diagnosis and treatment of autism and associated psychiatric comorbidities in young children.

Kathleen Campbell is a pediatrician in residency training at the University of Utah. The focus of her research is quality improvement of pediatric care and novel methods for autism screening and diagnosis. She completed her MD and Masters degree at Duke University School of Medicine. Kathleen also has worked on studies of brain development in children with autism and other developmental delays at the University of California, San Diego. She

earned a bachelors degree in molecular and cell biology with an emphasis in neurobiology at the University of California, Berkeley.

Qiang Qiu received his Bachelor's degree with first class honors in Computer Science in 2001, and his Master's degree in Computer Science in 2002, from National University of Singapore. He received his Ph.D. degree in Computer Science in 2013 from University of Maryland, College Park. During 2002–2007, he was a Senior Research Engineer at Institute for Infocomm Research, Singapore. He is currently an Assistant Research Professor at the Department of Electrical and Computer Engineering, Duke University. His research interests include computer vision and machine learning, specifically on face recognition, human activity recognition, image classification, and representation learning.

Steven Espinosa is a Senior IT Analyst working for Duke OIT and the Information Initiative at Duke (iiD). Steven is currently the lead developer for Autism & Beyond: A Study of young Childrens Mental Health (platform: iOS) as well as a technical advisor and developer for various efforts in the phone app, VR and AR space at Duke. After completing his BS in Computer Science at the University of South Carolina Steven has done software consulting, support, instruction and development in various contexts from the game industry, web and mobile. At Duke, Steven is committed to expanding the use of everyday devices for use in new and novel ways of doing research and medical care.

Samuel Marsan is currently a student in the doctoral program in clinical psychology at Duke University under the mentorship of Drs. Geraldine Dawson and Nancy Zucker. His interests include the assessment of construct validity for social interactions and somatic experiences through the Multitrait-Multimethod approach. His population of interest include persons at risk for autism and anorexia nervosa. Samuel earned a Bachelor of Science in psychology from The University of North Carolina at Chapel Hill in 2014.

Jeffrey P. Baker is director of the Trent Center for Bioethics, Humanities & History of Medicine. A professor of Pediatrics and History, he has practiced for over 25 years as a general pediatrician in Duke Childrens Primary Care with a focus on children with autism and special needs. Dr. Bakers historical work has also centered on child health. As the author of the book, The Machine in the Nursery (Johns Hopkins University Press, 1996), he is a leading authority on the history of neonatal medicine. He co-edited a 75th year anniversary history of the American Academy of Pediatrics, and has written historical articles related to pediatrics and vaccination. Dr. Baker also has directed the History of Medicine program within the Trent Center since 2006. He has directed the Duke Autism Clinic, and continues to engage actively with the Duke Center for Autism and Brain Development as its primary care liaison. On the national level, he chairs the committee advising the Pediatric History Center at the American Academy of Pediatrics and edits a historical monthly feature in the journal Pediatrics. He is also active within the American Association of the History of Medicine, where he currently serves as co-chair of the Program Committee.

Helen L. Egger is the Arnold Simon Professor and Chair of the Department of Child and Adolescent Psychiatry at New York University Langone Health (NYULH) and the Director of the NYULH Child Study Center. Previously, she was Chief of the Division of Child and Adolescent Psychiatry at Duke University Medical Center and Director of the Early Childhood Research Program in the Duke Center for Developmental Epidemiology. Dr. Egger is a child psychiatrist and epidemiologist whose research focuses on the developmental epidemiology and developmental neuroscience of early childhood psychiatric symptoms and disorders with a focus on anxiety, other emotional disorders, and autism. She also collaborates with pediatric mental health experts, computer engineers, software developers, and data scientists to develop apps for parents that use automatic computer vision, interactive design, and machine-learning analytics to screen and monitor young childrens behavior and social emotional development in their homes, schools, and communities. She co-led the team that released the first pediatric Apple ResearchKit app, Autism&Beyond, in the fall of 2015. Dr. Egger’s goal is to encode in apps the clinical and scientific knowledge within the NYU Langone Department of Child and Adolescent Psychiatry and the field of child psychiatry so as to extend clinical and scientific expertise to children and families who do not have access to this knowledge or to mental health care. Her innovation team is creating an early childhood digital platform with the release of an app on picky eating in summer 2018.

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 the Edmund T. Pratt, Jr. School 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, 0.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 and SIAM.

REFERENCES

[1]. Lord C, Risi S, Lambrecht L, Cook E, Leventhal B, DiLavore P, and Rutter M, “The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism,” Journal of Autism and Developmental Disorders, vol. 30, no. 3, pp. 205–23, 2000. [PubMed: 11055457]

[2]. Bryson S, Zwaigenbaum L, McDermott C, Rombough V, and Brian J, “The Autism Observation Scale for Infants: scale development and reliability data,” Journal of Autism and Developmental Disorders, vol. 38, no. 731–8, 2008.

[3]. Adrien JL, Faure M, Perrot A, Hameury L, Garreau B, Barthelemy C, and Sauvage D, “Autism and family home movies: preliminary findings,” Journal of Autism and Developmental Disorders, vol. 21, no. 1, pp. 43–9, 1991. [PubMed: 2037548]

[4]. Adrien JL, Perrot A, Sauvage D, Leddet I, Larmande C, Hameury L, and Barthelemy C, “Early symptoms in autism from family home movies. evaluation and comparison between 1st and 2nd year of life using I.B.S.E. scale,” Acta Paedopsychiatrica, vol. 55, no. 2, pp. 71–5, 1992. [PubMed: 1374996]

[5]. Werner E. and Dawson G, “Validation of the phenomenon of autistic regression using home videotapes,” Archives of General Psychiatry, vol. 62, no. 8, pp. 889–95, 2005. [PubMed: 16061766]

[6]. Mars A, Mauk J, and Dowrick P, “Symptoms of pervasive developmental disorders as observed in prediagnostic home videos of infants and toddlers,” Journal of Pediatrics, vol. 132, no. 3, pp. 500–4, 1998.

[7]. Osterling J. and Dawson G, “Early recognition of children with autism: a study of first birthday home videotapes,” Journal of Autism and Developmental Disorders, vol. 24, no. 3, pp. 247–57, 1994. [PubMed: 8050980]

[8]. Nadig A, Ozonoff S, Young G, Rozga A, Sigman M, and Rodgers S, “A prospective study of response to name in infants at risk for autism,” Archives of Pediatrics and Adolescent Medicine, vol. 161, no. 4, pp. 378–83, 2007. [PubMed: 17404135]

[9]. Elsabbagh M, Fernandes J, Webb S, Dawson G, Charman T, Johnson M, and British Austim Study of Infant Siblings Team, “Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood,” Biological Psychiatry, vol. 74, no. 3, pp. 189–94, 2013. [PubMed: 23374640]

[10]. Zwaigenbaum L, Bryson S, Rodgers T, Roberts W, Brian J, and Szatmari P, “Behavioral manifestations of autism in the first year of life,” International Journal of Developmental Neuroscience, vol. 23, no. 2, pp. 143–52, 2005. [PubMed: 15749241]

[11]. Ozonoff S, Losif A, Baguio F, Cook I, Hill M, Hutman T, Rodgers S, Rozga A, Sangha S, Sigman M, Steinfield M, and Young G, “A prospective study of the emergence of early behavioral signs of autism,” Journal of the American Acadameny of Child and Adolescent Psychiatry, vol. 49, no. 3, pp. 256–66, 2010.

[12]. Chawarska K, Macari S, and Shic F, “Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders,” Biological Psychiatry, vol. 74, no. 3, pp. 195–203, 2013. [PubMed: 23313640]

[13]. Constantino J, Kennon-McGill S, Weichselbaum C, Marrus N, Haider A, Glowinski A, Klaiman C, Klin A, and Jones W, “Infant viewing of social scenes is under genetic control and is atypical in autism,” Nature, vol. 547, pp. 340–4, 2017. [PubMed: 28700580]

[14]. Falck-Ytter T, Bölte S, and Gredebäck G, “Eye tracking in early autism research,” Journal of Neurodevelopmental Disorders, vol. 5, no. 28, pp. 1–13, 2013. [PubMed: 23402354]

[15]. Rehg J, Abowd G, Rozga A, Romero M, Clements M, Sclaroff S, Essa I, Ousley O, Yin L, Chanho K, Rao H, Kim J, Presti L, Jianming Z, Lantsman D, Bidwell J, and Zhefan Y, “Decoding children’s social behavior,” in IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3414–21.

[16]. Egmose I, Varni G, Cordes K, Smith-Nielsen J, Væver S, Køppe S, Cohen D, and Chetouani M, “Relations between automatically extracted motion features and the quality of motherinfant interactions at 4 and 13 months,” Frontiers in Psychology, vol. 8, no. 2178, pp. 1–16, 2017. [PubMed: 28197108]

[17]. Hashemi J, Tepper M, Spina T, Esler A, Morellas V, Papanikolopoulos N, Egger H, Dawson G, and Sapiro G, “Computer vision tools for low-cost and non-invasive measurement of autism-related behaviors in infants,” Autism Research and Treatment, 2014.

[18]. Martin K, Hammal Z, Ren G, Cohn J, Cassell J, Ogihara M, Britton J, Gutierrez A, and Messinger D, “Objective measurement of head movement differences in children with and without autism spectrum disorder,” Molecular Autism, vol. 9, no. 14, pp. 1–10, 2018. [PubMed: 29321841]

[19]. Leo M, Del Coco M, Carcagnìı P, Distante C, Bernava M, Pioggia G, and Palestra G, “Automatic emotion recognition in robot-children interaction for asd treatment,” in IEEE International Conference on Computer Vision Workshops, 2015.

[20]. Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, Centers for Disease Control and Prevention (CDC), “Prevalence of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network, 11 sites, United States, 2010,” Morbidity and Mortality Weekly Report Surveillance Summaries, vol. 63, no. 2, pp. 1–21, 2014.

[21]. Pickles A, Le Couteur A, Leadbitter K, Salomone E, Cole-Fletcher R, Tobin H, Gammer I, Lowry J, Vamvakas G, Byford S, Aldred C, Slonims V, McConachie H, Howlin P, Parr J, Charman T, and Green J, “Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial,” Lancet, vol. 388, no. 10059, pp. 2501–9, 2016. [PubMed: 27793431]

[22]. Jones EJH, Dawson G, Kelly J, Estes A, and Jane Webb S, “Parent-delivered early intervention in infants at risk for asd: Effects on electrophysiological and habituation measures of social attention,” Autism, vol. 10, no. 5, pp. 961–72, 2017.

[23]. Webb SJ, Jones EJ, Kelly J, and Dawson G, “The motivation for very early intervention for infants at high risk for autism spectrum disorders,” International Journal of Speech-Language Pathology, vol. 16, no. 1, pp. 36–42, 2014. [PubMed: 24410019]

[24]. Zwaigenbaum L, Bauman M, Stone W, Yirmiya N, Estes A, Hansen R, McPartland J, Natowicz M, Choueiri R, Fein D, Kasari C, Pierce K, Buie T, Carter A, Davis P, Granspeesheh D, Mailloux Z, Newschaffer C, Robins D, Roley S, Wagner S, and Wetherby A, “Early identification of autism spectrum disorder: Recommendations for practice and research,” Pediatrics, vol. 136, no. 1, pp. 10–40, 2015. [PubMed: 26055846]

[25]. Dawson G, Toth K, Abbot R, Osterling J, Munson J, Estes A, and Liaw J, “Early social attention impairments in autism: social orienting, joint attention, and attention to distress,” Developmental Psychology, vol. 40, no. 2, pp. 271–83, 2004. [PubMed: 14979766]

[26]. Egger H, Dawson G, Hashemi J, Carpenter K, Espinosa S, Campbell K, Brotkin S, Shaich-Borg J, Qiu Q, Tepper M, Baker J, Bloomfield R, and Sapiro G, “Automatic emotion and attention analysis of young children at home: A ResearchKit autism feasibility study,” npj Digital Medicine, vol. 1, no. 1, p. 20, 2018. [PubMed: 31304303]

[27]. Hashemi J, Campbell K, Carpenter K, Harris A, Qiu Q, Tepper M, Espinosa S, Shaich-Borg J, Marsan S, Calderbank R, Baker J, Egger H, Dawson G, and Sapiro G, “A scalable app for measuring autism risk behaviors in young children: A technical validity and feasibility study,” in EAI International Conference on Wireless Mobile Communication and Healthcare, 2015, pp. 23–7.

[28]. Robins D, Casagrande K, Barton M, Chen C, Dumont-Mathieu T, and Fein D, “Validation of the Modified Checklist for Autism in Toddlers, revised with follow-up M-CHAT-R/F,” Pediatrics, vol. 133, no. 1, pp. 37–45, 2015.

[29]. Luyster R, Gotham K, Guthrie W, Coffing M, Petrak R, Pierce K, Biship S, Esler A, Hus V, Rosalind O, Richler J, Risi S, and Lord C, “The Autism Diagnostic Observation Schedule – Toddler Module: a new module of standarized diagnostic measure for autism spectrum disorders,” Journal of Autism and Developmental Disorders, vol. 39, no. 9, pp. 1305–20, 2010.

[30]. De la Torre F, Chu W, Xiong X, Vicente F, Ding X, and Cohn J, “IntraFace,” in IEEE International Conference on Automatic Face and Gesture Recognition, 2015.

[31]. Manna S, Ruddock K, and Wooding D, “Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images,” Spatial Vision, vol. 9, no. 3, pp. 363–86, 1995. [PubMed: 8962841]

[32]. Li Y, Fathi A, and Rehg J, “Learning to predict gaze in egocentric video,” in IEEE International Conference on Computer Vision, 2013.

[33]. Hashemi J, Qiu Q, and Sapiro G, “Cross-modality pose-invariant facial expression,” in IEEE International Conference on Image Processing, 2015, pp. 4007–11.

[34]. Yin L, Wei X, Wang J, and Rosato M, “A 3D facial expression database for facial behavior research,” in IEEE Conference on Automatic Face and Gesture Recogition, 2006, pp. 211–216.

[35]. Ojala T, Pietikainen M, and Maenpaa T, “Multiresolution gray scale and rotation invariant texture classification with local binary patterns,” IEEE Transcations on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–87, 2002.

[36]. Chang C. and Lin C, “LIBSVM: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1–27, 2011.

[37]. Lucey P, Cohn J, Kanade T, Saragih J, Ambadar Z, and Matthews I, “The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression,” in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 94–101.

[38]. Oster H, “Baby FACS: Facial action coding system for infants and young children,” 2003, New York University.

[39]. Nodlus. [Online]. Available: http://www.noldus.com/human-behavior-research/products/the-observer-xt

[40]. Cicchetti D, “Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology,” Psychological Assessment, vol. 6, no. 4, pp. 284–90, 1994.

[41]. Shrout P. and Fleiss J, “Intraclass correlations: Uses in assessing rater reliability,” Psychological Bulletin, vol. 86, no. 2, pp. 420–8, 1979. [PubMed: 18839484]

[42]. Yutaka S, “The truth of the f-measure,” Teach Tutor Mater, 2007.

[43]. R Core Team R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[44]. Gamer M, Lemon J, Fellows I, and Singh P, irr: Various coefficients of interrater reliability and agreement, 0th ed., 2012. [Online]. Available: https://CRAN.R-project.org/package=irr

[45]. Sing T, Sander O, Beerenwinkel N, and Lengauer T, “ROCR: visualizing classifier performance in R,” Bionformatics, vol. 21, no. 20, p. 7881, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de

[46]. Campbell K, Carpenter K, Hashemi J, Espinosa S, Marsan S, Schaich Borg J, Chang Z, Qiu Q, Alder E, Vermeer S, Tepper M, Egger H, Baker J, Sapiro G, and Dawson G, “Computer vision analysis captures atypical attention in toddlers with autism,” Autism, 2018.

[47]. Guha T, Yang Z, Grossman R, and Narayanan S, “A computational study of expressive facial dynamics in children with autism,” IEEE Transcations on Affective Computing, vol. 9, no. 1, pp. 14–20, 2018. [PubMed: 29963280]

[48]. Simon T, Joo H, Matthew I, and Sheikh Y, “Hand keypoint detection in single images using multiview bootstrapping,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4645–53.

[49]. Liu N, Han J, Zhang D, Wen S, and Liu T, “Predicting eye fixations using convolutional neural networks,” in IEEE International Conference on Computer Vision, 2015.

[50]. Papoutsaki A, Sangkloy P, Laskey J, Daskalova N, Huang J, and Hays J, “Webgazer: Scalable webcam eye tracking using user interactions,” in International Joint Conference on Artificial Intelligence, 2016.

[51]. Krafka K, Khosla A, Kellnhofer P, Kannan H, Bhandarkar S, Matusik W, and Torralba A, “Eye tracking for everyone,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[52]. Funes-Mora K. and Odobez J, “Gaze estimation in the 3D space using RGB-D sensors,” International Journal of Computer Vision, vol. 118, no. 2, pp. 194–216, 2016.

[53]. Cazzato D, Leo M, and Distante C, “An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation,” Sensors, vol. 14, no. 5, pp. 8363–79, 2014. [PubMed: 24824369]

[54]. Tseng P, Cameron I, Pari G, Reynolds J, Munoz D, and Itti L, “High-throughput classification of clinical populations from natural viewing eye movements,” Journal of Neurology, vol. 260, no. 1, pp. 275–84, 2013. [PubMed: 22926163]

[55]. Wang S, Jiang M, XM D., Laugeson E, Kennedy D, Adolphs R, and Zhao Q, “Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking,” Neuron, vol. 88, no. 3, pp. 604–16, 2015. [PubMed: 26593094]

[56]. Jiang M. and Zhao Q, “Learning visual attention to identify people with autism spectrum disorder,” in IEEE International Conference on Computer Vision, 2017.

[57]. Liu W, Li M, and L Y, “Identifying children with autism spectrum disorder based on their face processing abnormality: a machine learning framework,” Autism Research, 2016.

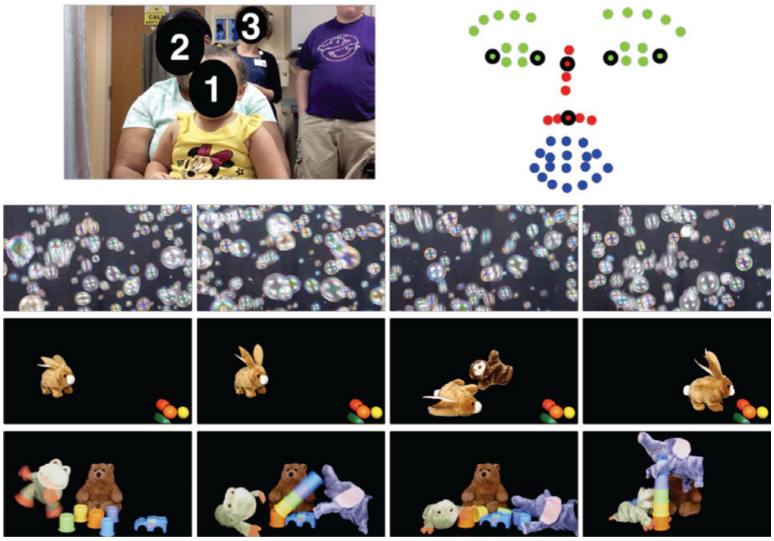

Fig. 1:

Screenshot of the recorded video from the front facing tablet's camera and example of automatic facial landmarking are shown in first row. In this screenshot, the child (1) is sitting on the caregiver's (2) lap, while the practitioner (3) is standing behind. All six outlined automatically detected landmarks (in black) are used for face pre-processing, while the lowest nose and the two outer eye landmarks are used to track head movement. Screenshots of frames from the movie stimuli being presented are shown in the remaining rows. These are Bubbles, Bunny, and Puppets, respectively.

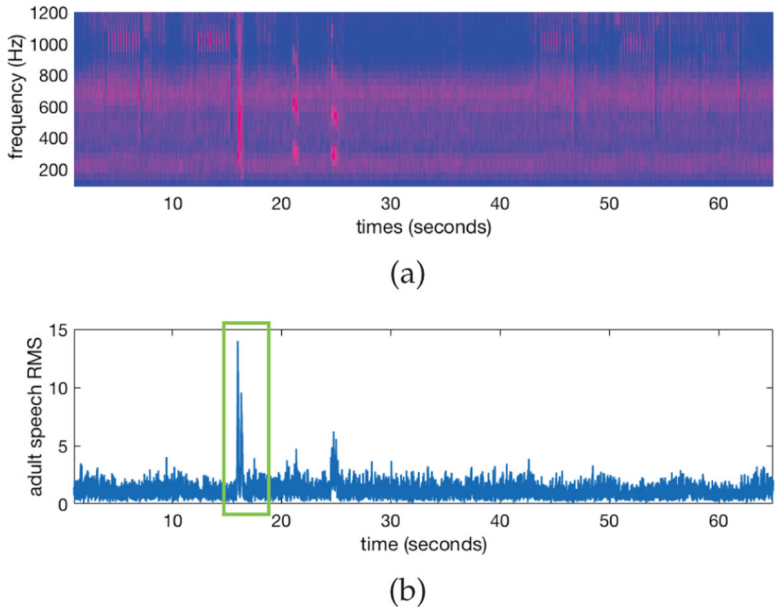

Fig. 2:

Audio is analyzed to detemine the exact time point the practitioner said the child's name during a name-call. The power spectrum density (psd) of the recorded audio signal (2a) contains audio from the movie stimuli (predominantly music) and instances of vocalizations. Root mean squared (RMS) values of the audio signal (2b) provide quantification of audio signals at each time point, and are used to detect a name-call prompt. Knowing that practitioner was asked to prompt a name-call at 15 seconds into the stimuli, in this example we are able to focus on speech around the time point (green box) and detect the exact time point when maximum speech occurred.

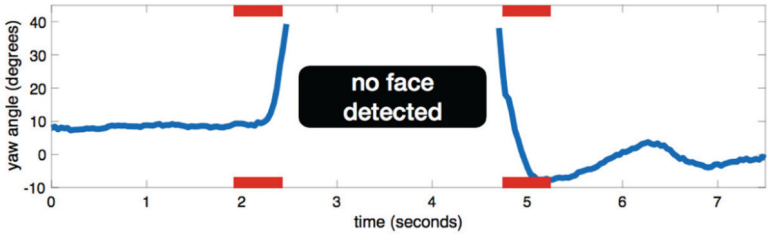

Fig. 3:

Example of a head turn using the automatic method. To differentiate a head turn from a face occlusion, we determine if the child is performing a head turning motion before and after the face is lost or when its exhibiting a yaw pose with large magnitude. The red bars represent the half-second windows used to determine if the child is exhibiting a head turning motion before and after the face is lost (by the camera) or when its exhibiting a yaw pose with large magnitude.

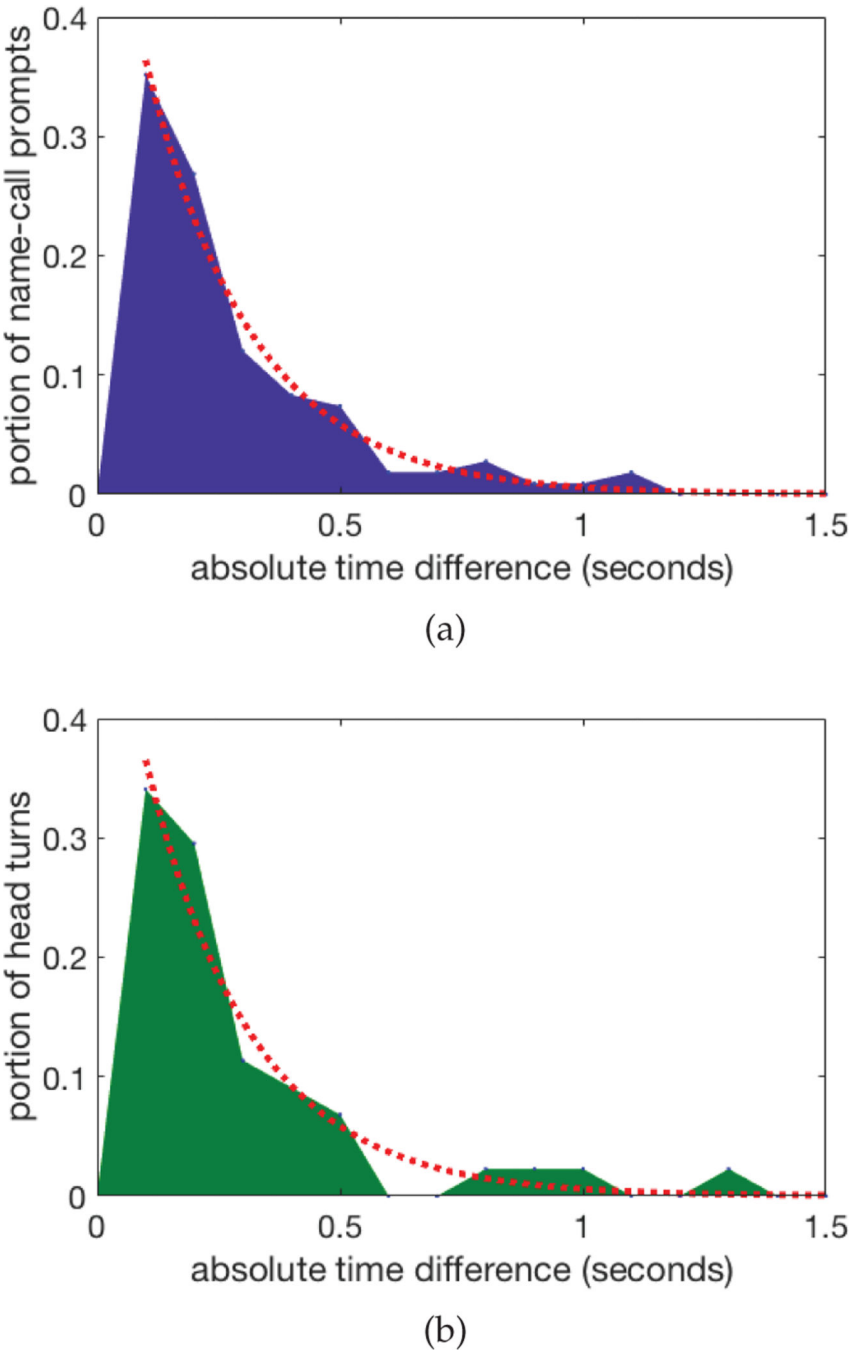

Fig. 4:

Area plots of absolute time difference between automatic methods and hand labeled data for name-call prompt detection (4a) and for head turn detection (4b). Fitted exponential curves are shown in the dotted red lines.

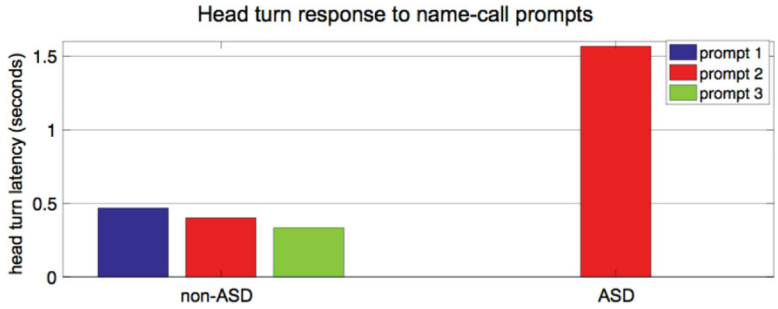

Fig. 5:
Examples of responses from a non-ASD and ASD toddler to name-call.

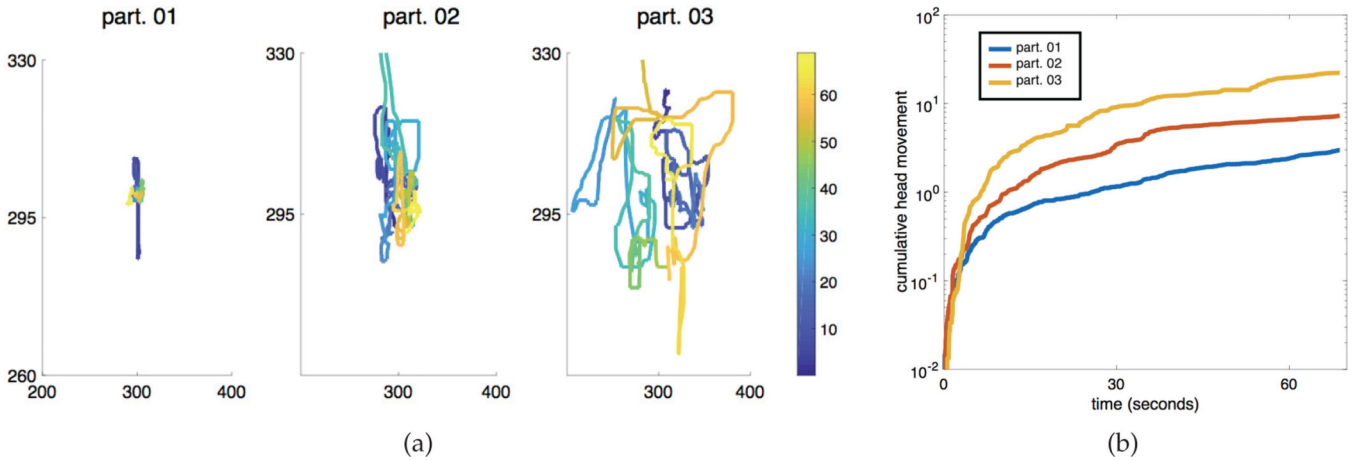

Fig. 6:

Examples of the head movement of three participants during the Puppets stimuli. 6a shows the head movements of the participants, where the axis represent pixel coordinates in the video recording. The lines are color-coded with respect to time, where the colorbar on the right represents time (seconds) in the movie stimuli. A log-plot of the cumulative head movement (as a simple quantifying measure) for all three participants is shown in 6b. Figure is best viewed in color.

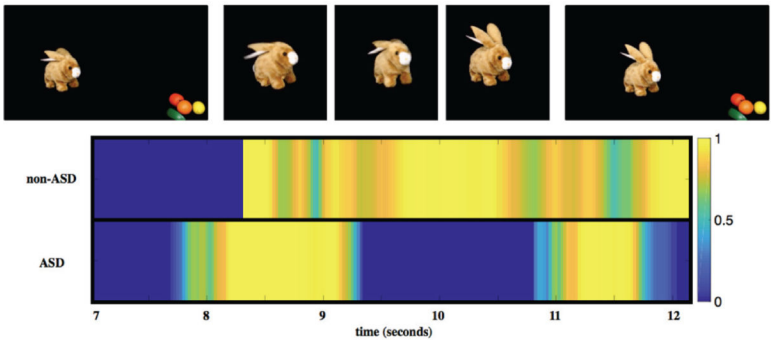

Fig. 7:

Probability scores of expressing Happy for a non-ASD (top) and ASD (bottom) toddler reacting to a scene during the Bunny movie stimuli. Screenshots of the stimuli are shown in the first row; in this scene in the movie the bunny is jumping and then stopping and making noises while moving its ears and nose. The colorbar on the right indicates probability scores of expressing Happy. Figure is best viewed in color.

TABLE 1:

Demographics table reported in means (standard deviations) or counts (percentage)

	non-ASD	ASD
Total participants	18	15
Males	16 (89%)	13 (87%)
Age in months	26.4 (std 4.6)	25.5 (std 3.8)
ADOS-T Total	-	17.7 (std 5.3)

TABLE 2:

Number of video recordings

Group	Bubbles	Bunny	Puppets	Total
non-ASD	18	18	18	54
ASD	15	15	15	45

TABLE 3:

Inter-rater reliability results of automatic codings to an expert human rater. Reported are ICC (95% CI)

Coding	ALL	non-ASD	ASD
Engagement	**0.85** (0.77–0.89)	0.89 (0.81–0.94)	0.81 (0.69–0.90)
Name-call response	**0.84** (0.67–0.91)	0.86 (0.72–0.93)	0.80 (0.66–0.89)
Total time exhibiting Happy	**0.90** (0.85–0.93)	0.89 (0.81–0.94)	0.90 (0.84–0.95)

IEEE Trans Affect Comput. Author manuscript; available in PMC 2022 April 08.

TABLE 4:

To validate the automatic emotion classification and compare it with an expert human rater, precision, recall, and F1-scores are reported across groups

	ALL (136,450 frames)				non-ASD (78,299 frames)				ASD (58,151 frames)			
Emotion	**Frequency**	**Precision**	**Recall**	**F1**	**Frequency**	**Precision**	**Recall**	**F1**	**Frequency**	**Precision**	**Recall**	**F1**
Happy	0.20	0.81	0.70	0.75	0.20	0.85	0.68	0.76	0.19	0.75	0.74	0.74
Other	0.80	0.91	0.95	0.93	0.80	0.88	0.96	0.92	0.81	0.94	0.94	0.94
Overall	**1**	**0.89**	**0.90**	**0.89**	**1**	**0.87**	**0.90**	**0.89**	**1**	**0.90**	**0.90**	**0.90**

The reported overall scores are weighted averages based on frequency.