# A computer vision approach for the assessment of autism-related behavioral markers

Jordan Hashemi
Mariano Tepper
Guillermo Sapiro
ECE - University of Minnesota

Thiago Vallin Spina
Institute of Computing
University of Campinas

Vassilios Morellas
Nikolaos Papanikolopoulos
CSE - University of Minnesota

Amy Esler
Department of Pediatrics
University of Minnesota

***Abstract*—The early detection of developmental disorders is key to child outcome, allowing interventions to be initiated that promote development and improve prognosis. Research on autism spectrum disorder (ASD) suggests behavioral markers can be observed late in the first year of life. Many of these studies involved extensive frame-by-frame video observation and analysis of a child's natural behavior. Although non-intrusive, these methods are extremely time-intensive and require a high level of observer training; thus, they are impractical for clinical purposes. Diagnostic measures for ASD are available for infants but are only accurate when used by specialists experienced in early diagnosis. This work is a first milestone in a long-term multidisciplinary project that aims at helping clinicians and general practitioners accomplish this early detection/measurement task automatically. We focus on providing computer vision tools to measure and identify ASD behavioral markers based on components of the Autism Observation Scale for Infants (AOSI). In particular, we develop algorithms to measure three critical AOSI activities that assess visual attention. We augment these AOSI activities with an additional test that analyzes asymmetrical patterns in unsupported gait. The first set of algorithms involves assessing head motion by facial feature tracking, while the gait analysis relies on joint foreground segmentation and 2D body pose estimation in video. We show results that provide insightful knowledge to augment the clinician's behavioral observations obtained from real in-clinic assessments.**

## I. Introduction

The analysis of children's natural behavior is of key importance for the early detection of developmental disorders such as autism spectrum disorder (ASD). For example, several studies have revealed behaviors indicative of ASD in early home videos of children that were later diagnosed with ASD (see [1] and references therein). These studies involved video-recording these environments and then analyzing the videos a posteriori, using frame-by-frame viewing by an observer who typically trains for several weeks to achieve inter-rater reliability. Of course, hours and hours of labor are required, making such analyses impractical for clinical settings. While clinical tools for early diagnosis of ASD are available, they require administration and interpretation by specialists in early diagnosis of ASD. Most families lack easy access to specialists in ASD; for example, the wait list for an evaluation at the leading ASD Clinic at our university is 6 months for children age 4 and under. There is a need for automatic and quantitative analysis tools that can be used by general practitioners in child development, and in general environments, to identify children at-risk for ASD and other developmental disorders.

As a first milestone in this long-term goal, this work focuses on providing computer vision tools for aiding in-clinic early diagnosis of ASD. Although much is unknown about the underlying causes of ASD, it is characterized by abnormalities in social interactions and communication and the presence of restricted, repetitive behaviors [1]. Zwaigenbaum et al. [1] argue that children with ASD exhibit several specific behavioral markers as early as in the first year of life. These markers appear, among others, in activities involving visual attention, often expressed as dificulties in disengagement and shifting of attention [2]. Many children also have atypical motor patterns, such as asymmetric gait or toe walking [3].

Despite this evidence, the average age of ASD diagnosis in the US is 5 years [4]. Recently, much research and clinical trials have focused on early diagnosis to allow for early intensive intervention. Early intervention, initiated in preschool and sustained for at least 2 years, can substantially improve child outcomes (e.g., [5]). Detecting ASD risk and starting interventions before the full set of behavioral symptoms appears has an even greater impact, preventing difficult behaviors and delayed developmental trajectories from taking hold [5]. Early diagnosis is achieved by following a comprehensive battery of measurable tests and parent interviews, with the goal of detecting behavioral symptoms consistent with ASD. In this work, we develop semi-automatic computer vision video analysis techniques to aid in such diagnostic assessment.

These tools aid the clinician in the diagnosis task by providing accurate and objective measurements. In addition, and particularly for research, automatic analysis will permit to effortlessly analyze vast amounts of naturally recorded videos, opening the door for data mining towards the improvement of current assessment protocols and the discovery of new behavioral features. This project is being developed by a multidisciplinary group bringing together professionals from psychology, computer vision, and machine learning. As opposed to other research projects, e.g., [6], where artificial setups are used, one of our main goals is to provide non-intrusive capturing systems that do not induce behavioral modification in the children. In other words, hardware must not constraint the testing environment: the clinician is free to adjust testing conditions as needed, and children are not asked to wear any type of sensors or perform any non-natural tasks.

The results in this paper are from actual clinical recordings,

J. Hashemi, T. V. Spina, and M. Tepper equally contributed to this work.

Work partially supported by NSF Grants 1039741 & 1028076, and CAPES (BEX 1018/11-6) and FAPESP (2011/01434-9) PhD scholarships from Brazil.

in which the at-risk infant/toddler is tested by an experienced clinician following the Autism Observation Scale for Infants (AOSI) [7] and a standard battery of developmental and ASD assessment measures. The AOSI involves a set of semi-structured activities that provide an interactive context in which the examiner engages the infant in play, while conducting a set of systematic presses to elicit specific child behaviors. In our clinical setup, we use two low-cost GoPro Hero HD color cameras (resolution: 1080p at 30 fps), one placed on the clinician's table (e.g., Fig. 2) and one in a corner of the room (Fig. 5); the displayed images are downsampled, blurred, and/or partially blocked to preserve anonymity (processing was done on the original videos).[1]

In this work we address four fundamental behaviors:

**Sharing Interest (AOSI).** It is described as the "ability to use eyes to reference and share interest in an object or event with another person" [7]. Although this behavior is evaluated throughout the AOSI, it can be specifically assessed from a ball playing activity: a ball is rolled on the table towards the infant after engaging his/her attention. After receiving the ball, the child's ability to acknowledge the involvement of another person in the gameplay by looking to her is assessed.

**Visual Tracking (AOSI).** It represents the "ability to visually follow a moving object laterally across the midline" [7]. To evaluate it, the following activity is performed: (1) a rattle or other noisy toy is used to engage the infant's attention, (2) the rattle is positioned to one side of the infant, and (3) the rattle is then moved silently at eye level across the midline to the other side. The ability to track laterally the rattle is assessed.

**Disengagement of Attention (AOSI).** It is characterized as the "ability to disengage and move eyes/attention from one of two competing visual stimuli" [7]. The corresponding activity consists of (1) shaking a noisy toy to one side of the infant until his/her attention is engaged, and (2) then shaking a second noisy toy on the opposite side, while continuing to shake the first object. The ability to disengage and move his/her eyes/attention from the first to the second object is assessed.

**Atypical Motor Behavior (full session).** It is portrayed in the AOSI as the "presence of developmentally atypical gait, locomotion, motor mannerisms/postures or repetitive motor behaviours" [7]. There is no specific activity for assessing motor patterns. The clinician performs a holistic evaluation of the behaviors whenever they occur throughout the full session.

We present video analysis tools for assessing these behavioral patterns. Both the application and the work with such specific population of infants and toddlers are unique in the vision community. The first three behaviors will be covered in Section II, by tracking simple facial features and estimating the head movements from them. The last behavior is treated in Section III using a joint segmentation/pose estimation algorithm. Some final remarks are provided in Section IV.

## II. Assessing Visual Attention

Through the development of the AOSI, Zwaigenbaum et al. [1] identified multiple behavioral markers for early detection of ASD. We focus on three of these, namely sharing interest, visual tracking, and disengagement of attention. The AOSI states specific guidelines on how to evaluate these behavioral markers from their corresponding activities.

**Sharing Interest.** We focus on a ball play task from the AOSI as a measure of shared interest. The clinician analyzes how the child uses his/her eye gaze to share interest and enjoyment with him, by looking from the ball to the clinician. Infrequent or limited looking to faces is an early ASD risk sign [1,7].

**Visual Tracking.** The clinician evaluates how well the infant tracks the moving object. Infants with ASD usually exhibit discontinuous and/or a noticeably delayed tracking [7].

**Disengagement of Attention.** The clinician assesses the child's ability to shift attention away from one object when another is presented. A delayed response is an ASD risk sign [2].

Our focus is to create a semi-automatic, non-intrusive, and accurate tool for aiding in the evaluations of these three behaviors related to ASD risk. There have been recent advancements in creating semi-automatic tools to analyze some ASD risk signs, e.g., [6]; however they all involve controlling the child's environment. There have also been a lot of recent great accomplishments in robust head pose estimation of adults [8]–[10]. However, Wen et al. [11] showed that tools based on adults are not effective when used on infants. We use 4 facial features (eyes, left ear, and nose) to estimate the child's yaw (left to right) motion, and 3 facial features (left eye, left ear, and nose) to estimate pitch (up and down) motion.

### A. Tracking and Validating Facial Features

The large variability of the data, and the non-optimal positioning of the camera, calls for very simple and robust features and algorithms as described next. We assume that, in the first frame, we have bounding boxes of three facial features: the left ear, left eye, and nose. To track these three facial features, and following a scheme loosely based on [12], we use dense motion estimaton coupled with a validation step that employes an offline-trained facial feature detector. The dense motion estimator tracks the features with high accuracy in most cases, but when the child's head moves quickly, illumination changes can sometimes cause the tracker to lag behind the features. Thus we validate the output of the tracker using facial feature detectors in every frame.

To validate the features we train left eye, right eye, left ear, and nose detectors based on the method proposed in [13] (see also [14]). Our method uses multiscale Histograms of Orientated Gradients (HOG) as descriptors to represent each facial feature, and then classifies these descriptors using a Support Vector Machine. As positive training samples, we use hand labeled facial patches from children in our experimental environment. As negative training samples, we extract random patches from around multiple children's faces.

For each frame, search areas for the facial feature detectors are defined around the bounding boxes given by the tracker. Since the left eye, left ear, and nose are present in every frame for the given camera position, we impose a lenient classifier threshold and geometrical constraints (e.g., the left eye must be higher and to the left of the nose). The tracker's bounding boxes are validated if their centers are within the bounding

[1]Approval for this study was obtained from the Institutional Review Board at the University of Minnesota.

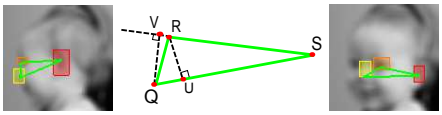


Fig. 1. Examples of the triangle created by the left ear, left eye, and nose. The leftmost and rightmost images depict the triangle when the infant is looking right and more towards the camera, respectively. The middle image shows the points used for calculating $\widehat{\text{yaw}}$.

boxes returned by the detectors; however, if the tracker's centers are outside of the detector's bounding boxes for two consecutive frames, then the corresponding bounding box for the tracker is reset to a new location within the detector's bounding box. Determining the presence of the right eye aids in the estimation of the yaw motion. The search area for the right eye, which is not tracked since it appears and disappears constantly, is based on the locations of the detected left eye and nose.

### *B. Yaw and Pitch Motion Estimation from Facial Features*

To analyze the child's reactions in the visual attention activities, we estimate the changes of two head pose angles: yaw and pitch. For the Visual Tracking and Disengagement of Attention activities, which involve lateral motions, we focus on the changes of the yaw angle; conversely, in the Sharing Interest activity, we focus on the changes of the pitch angle.

The child's face is predominantly in a profile view for the Sharing Interest activity. As a way to provide an accurate motion estimation of the pitch angle we cumulatively sum the vertical coordinate changes of the left eye and nose with respect to the left ear across a period of 2 frames. We expect a positive sum when the child is looking up and a negative sum when the child is looking down, the magnitude representing how much the child is looking up or down.

For estimating the yaw angle motion in the Visual Tracking and Disengagement of Attention activities, we calculate two ratios based on the triangle created by the left ear, left eye, and nose (Fig. 1); we also use information about the presence of the right eye. Let $Q$, $R$, and $S$ denote the locations of the nose, left eye, and left ear, respectively. For the first ratio $r_{\text{NoseToEye}}$, we project $R$ into the line defined by $QS$, thus defining the point $U$; we then define $r_{\text{NoseToEye}} = |US|/|QS|$, where $|\cdot|$ is the Euclidian distance. For the second ratio we project $Q$ into the line defined by $RS$, defining $r_{\text{EyeToEar}} = |VR|/|RS|$.

The two ratios $r_{\text{EyeToEar}}$ and $r_{\text{NoseToEye}}$ are inversely proportional. Looking at Fig. 1 we can observe that when the face is looking in profile view, $r_{\text{EyeToEar}}$ will be large and $r_{\text{NoseToEye}}$ will be low; conversely when the face is in frontal view (looking more towards the camera). To combine these two ratios into one value, we calculate the normalized difference between them, $\widehat{\text{yaw}} = \frac{r_{\text{EyeToEar}} - r_{\text{NoseToEye}}}{r_{\text{EyeToEar}} + r_{\text{NoseToEye}}}$. Thus, as the child is looking to his/her left, $\widehat{\text{yaw}}$ goes to -1; and as the child is looking to his/her right, $\widehat{\text{yaw}}$ goes to 1. The presence of the right eye further verifies that the infant is looking left.

### *C. Experimental Results*

Our sample includes 3 ASD-risk and 3 non-ASD-risk infants (ages 6 - 15 months). To initialize our algorithm we defined a bounding box of the left ear, left eye, and nose in the first frame only.[2] We marked the playing objects by hand, although these can be done automatically from prior knowledge of their features. We present two examples for each of the three AOSI tasks. In all cases, our algorithm properly tracks the infants' facial features and estimates the yaw or pitch motions;

Fig. 2 shows examples of our results for the Sharing Interest activity. Our algorithm confirms that both infants were able to track the ball as it was being rolled to them, and that they both looked back up at the clinician after receiving the ball, providing evidence of shared interest. Our algorithm is of course able to record quantitative measures such as how long it takes for the infant to look from the ball back up to the clinician, such types of measurements having merit both for detecting an ASD risk marker and for the study of large populations towards the discovery of new markers.

Fig. 3 shows examples of our results for the Disengagement of Attention activity. Both infants were able to recognize the second object after it was presented; however, there was a noticeable difference in speed. Our algorithm shows that the first infant identified the second object within 0.6s of it being presented; while it took the second infant nearly 1s to identify the second object. Following the AOSI, such delay is scored as an indication of ASD risk. On the AOSI, a trial is considered "passed" if the child looks to the second object in less than 1s, considered "delayed" if the child looks after a 1-2s delay, and considered "stuck" if the child looks after more than 2s. The clinician makes a "live" judgment about this time frame or may look at videos of this task if available; automating this judgment may improve accuracy. Again, such automatic and quantitative measurements are critical for aiding diagnosis.

Fig. 4 shows examples of our results for the Visual Tracking activity. During this activity, we focus on the child's ability to smoothly track the object. From our results, the first infant smoothly tracked the moving object; while the second infant's motion was less smooth, a potential indicator of ASD risk.

Our results are clinically consistent, since both the automatic and expert assessments agree for the studied video segments. Note that the overall diagnosis takes into account several factors throughout the full session.

## III. Assessing Motor Patterns

According to Esposito et al. [3, and references therein], motor development has often been hypothesized as an early bio-marker of autism, and motor development disorders are considered one of the first signs which could precede social or linguistic abnormalities. Hence, it is important to find means of detecting and measuring these atypical motor patterns at a very early stage. Children diagnosed with autism may present asymmetric gait patterns when walking unsupportedly. In particular, these authors [3] have found that diagnosed toddlers often presented asymmetric arm positions (Fig. 6), according to the Eshkol-Wachman movement notation (EWMN) [15], in home videos filmed during the children's early life period. EWMN is essentially a 2D stickman that is manually adjusted to the child's body on each video frame and then analyzed.

[2]See all video results at http://baarc.cs.umn.edu/icdl-epirob/results.zip, in folder "visual_attention", username: icdl-epirob2012, password: reviewers.

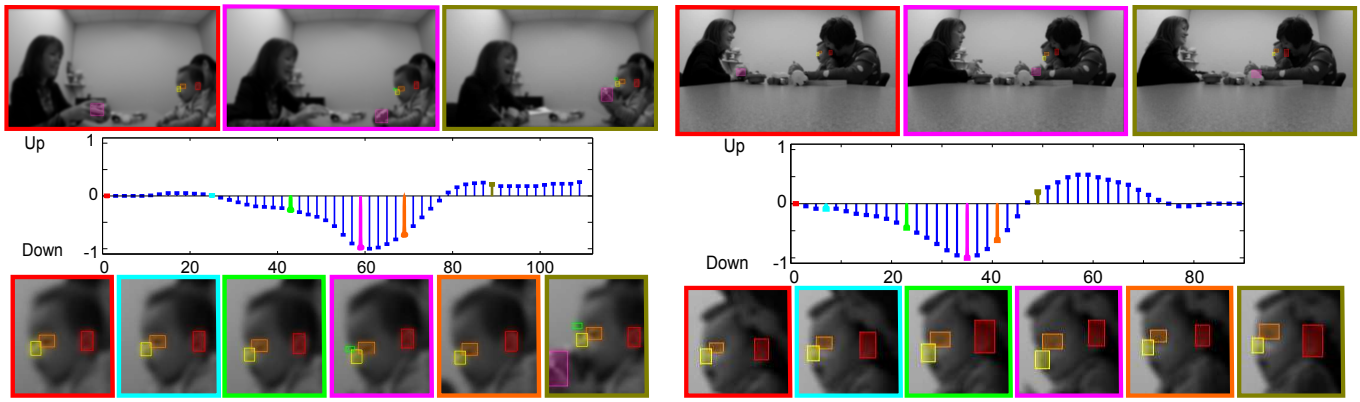


Fig. 2. Sharing Interest activity. **Top:** start of the activity, when the ball contacts the child, and when the child looks up at the clinician. **Middle:** changes in the pitch motion (y-axis) for every other frame (x-axis). **Bottom:** 6 examples (3 of which correspond to the 3 top scenes) of the infant's face during the administration. All face features are automatically detected and tracked. The coloring of the bounding box around the scenes and infant's faces corresponds to the pitch analysis in the respectively colored frame.

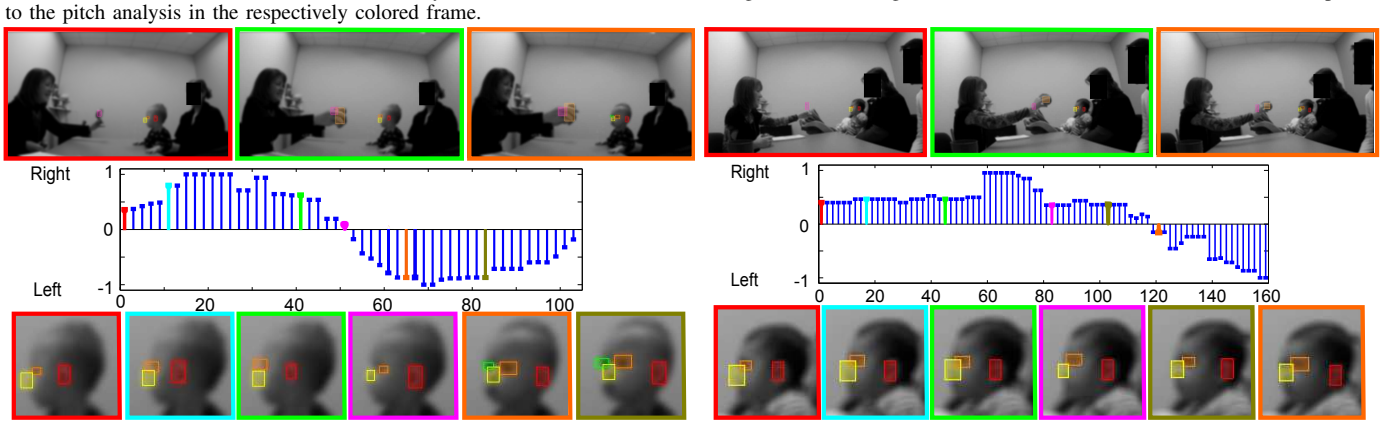


Fig. 3. Disengagement of Attention activity. **Top:** clinician holding one object, when the second object appears, and when the child recognizes the second object. **Middle:** changes in the yaw motion ($\widehat{\text{yaw}}$ values in the y-axis) for every other frame (x-axis). **Bottom:** 6 examples of the infant's face during the activity. Coloring is as in the previous figure.

Our goal is to semi-automate this task by estimating the 2D body pose of the toddlers in video segments in which they are walking unconstrained. This methodology is supported by the AOSI protocol, since the assessment of motor functionalities is done by visual inspection throughout the session [7].

Human body pose estimation is a complex and relatively well explored research topic in computer vision, e.g., [16,17], although it has been mostly restricted to adults, often in constrained scenarios, and not yet exploited in the application we address. We approach 2D human pose estimation by extending the Object Cloud Model (OCM) segmentation framework [18] to work with articulated structures and video data.[3]

The OCM is represented by a *fuzzy object* (*cloud image*) in which each pixel receives one out of three possible values: object, background, or uncertainty. The silhouette variations are captured by the uncertainty region, which represents the area where the real object's boundary is expected to be in a new test image (Fig. 5). OCM then treats the object detection task (locating the object of interest in an image) and delineation (defining the object's spatial extent) in a synergistic fashion. Namely, for each possible object position in an image (frame), OCM executes a delineation algorithm in the uncertainty region and evaluates if the resulting segmentation mask yields a maximum score for a given search criterion. This maximum should be reached when the uncertainty region is properly positioned over the real object's boundary. Ideally, if the uncertainty region is well adapted to the object's new silhouette and the delineation is successful, the object search is reduced to translating the model over the image.

When the object is composed of multiple correlated substructures, such as the parts of the human brain, a Cloud System Model (CSM) may be created by transforming each substructure into an OCM and taking into account the relative position between them during the search [18]. We consider the human body as the object of interest, divide it into each of its major structures (torso, head, arms, and legs), and connect those structures using a 2D stickman model to create a CSM in a given initial frame (figs. 5(a)-(b)). Then, the resulting CSM is used to automatically find the toddler's body frame-by-frame in the video segment (figs. 5(c)-(d)) and the corresponding stickman poses are used in our measurements.

### A. Foreground Segmentation Using the Cloud System Model

Formally, the CSM is a triple $C = (\mathcal{O}, A, F)$, composed of a set $\mathcal{O}$ of fuzzy objects $O_l$ (one OCM for each body part $l = 1, \ldots, m$), a delineation algorithm $A$, and a recognition functional $F$. A fuzzy object $O_l$ is a function that encodes the likelihood $O_l(x) \in [0, 1]$ of pixel $x$ belonging to body part $l$ (Fig. 5(b)). We require a single segmentation mask $L^t$, provided interactively at a given initial frame $I^t$, to

[3]The OCM extension was jointly developed with Alexandre X. Falcão.

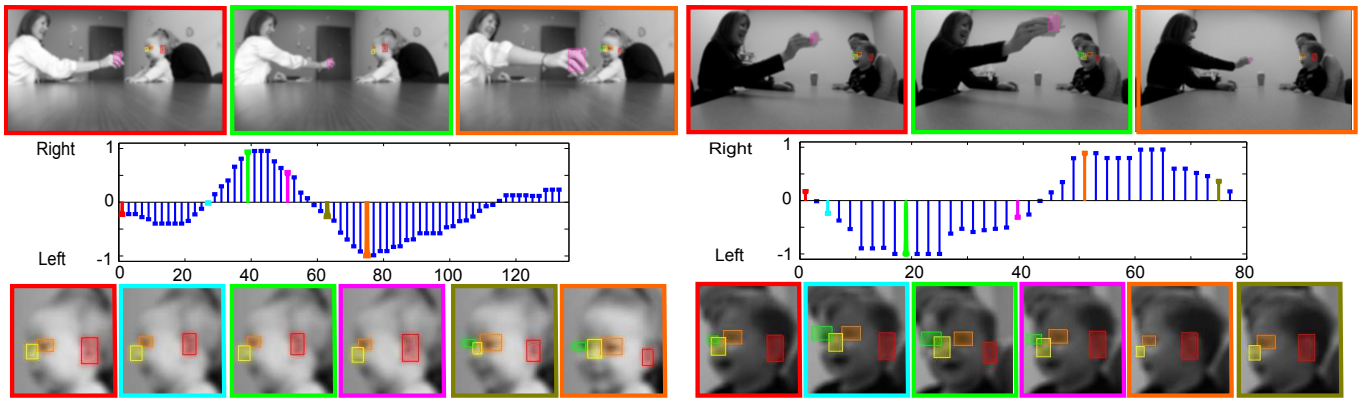


Fig. 4. Visual Tracking activity. **Top:** the clinician holding the object, when the object is at one extreme side (right or left), and when the object is at the other extreme side. See Fig. 3 for details on each panel.

compute $\mathcal{O}$. We construct $\mathcal{O}$ by applying an Euclidian distance transform to the border of each binary segmentation mask corresponding to one of the $m$ body parts in $L^t$. The distance transforms are further smoothed using a sigmoid function to control the size of the uncertainty region of each cloud image individually (figs. 5(a)-(b)).

Our delineation algorithm $A$ uses the rgaph-based framework of the Image Foresting Transform (IFT) [19]. By defining a suitable image-graph and connectivity function, delineation can be constrained to the uncertainty region of each OCM [18]. We fuse this with the work in [20] to achieve an accurate delineation in a test frame $I^{t'}$, $t' > t$ (Fig. 5(d)). Namely, we estimate the weights for the IFT image-graph by computing a rich image gradient that combines texture information from the test frame $I^{t'}$ with a shape-based constraint derived from the cloud image [18], and adds foreground color information obtained from local classifiers positioned and computed around the uncertainty region of each cloud [20].

Finally, our recognition functional $F$ (search criterion) takes into account the comparison of color histograms across frames to ouput a score for the delineation result during the object search. More precisely, a set of object and background color histograms are computed in local windows spread around the uncertainty region of each cloud image $O_l$, considering the quantized RGB colorspace (4 bins per channel). These histograms are (re)defined after each search delineation in frame $I^{t'}$, and use the pixels which were labeled by IFT as object or background, respectively. Then, the recognition functional $F$ for the current search position is the average $\chi^2$ distance between the object (background) histograms of frames $I^{t'}$ and $I^{t'-1}$, for all windows of each cloud $O_l$. Hence, the maximization of $F$ is turned to a minimization problem over the average $\chi^2$ distance. After finding the body in frame $I^{t'}$, all object and background histograms computed for the final position are stored and used for comparison when searching for the toddler's body in frame $I^{t'+1}$.

### B. Pose Search in Video Frames Using the CSM

We extend the CSM definition to include a relational model $\Theta$ that determines how the individual clouds are connected, as well as the possible relative angles between them. In this work, $\Theta$ is a 2D stickman representation in the form of a tree rooted at the torso (Fig. 5(b)). Each body part/OCM is

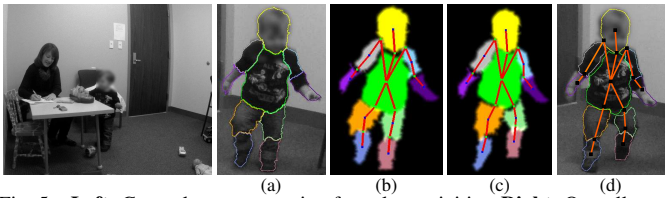


Fig. 5. **Left:** General scene capturing free play activities. **Right:** Overall segmentation and position tracking scheme. (a) Segmentation mask $L^t$ provided at an initial frame $t = 0$. (b) CSM computed from $L^t$ and the 2D stickman, used to connect the clouds corresponding to each body part. (c) Transformed CSM at frame 8. (d) Segmentation and final pose estimation.

connected by the main joint between them (elbow, shoulder, hip, and knee), which can be computed from the intersection of the clouds' main axes in the initial frame $I^t$. We also allow changes in scale for all clouds in order to cope with projective transformations (e.g., limb foreshortening).

The search for the toddler's body in a frame at time $t' > t$ is done in a hierarchical fashion starting from the torso and involves optimizing $\Theta$, by minimizing the recognition functional $F$ for each OCM. Note that only the torso (i.e., the root of the tree) is translated over the image, while the other body parts are carried along with it. The optimization of $\Theta$ may be done using any standard algorithm. We use $\Theta^{t'-1}$ as an initialization to find the new configuration $\Theta^{t'}$, and improve this guess by first propagating the delineation label $L^{t'-1}$ at frame $I^{t'-1}$ to the current frame $I^{t'}$ using dense optical flow. Being $L^*$ the result of the propagation of label $L^{t'-1}$ to frame $I^{t'}$, we evaluate the changes in rotation and scale for each body part $l = 1, 2, \ldots, m$ between $L^{t'-1}$ and $L^*$ and use them as a starting point to find the new state $\Theta^{t'}$ (figs. 5(c)-(d)). During this evaluation we do not allow sudden state changes by constraining the CSM by the stickman relational model (i.e., changes in the relative angle between body parts should be smaller than $15^o$). However, we do permit that the joints move with partial freedom, proportionally to their body part's optical flow, to achieve finer adjustments.

### C. Experimental Results

We tested our algorithm in clips in which the entire body of the child can be seen [3]. We illustrate two such examples in Fig. 6 (with groundtruths).[4] These segments are typically

[4]See video results at the link from footnote 2, p. 3, folder "motor_pattern".

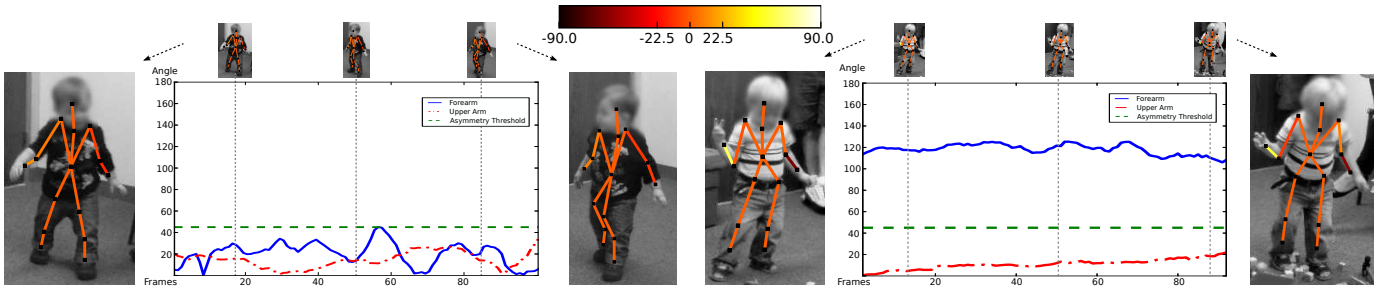


Fig. 6. Pose estimation performed in two video segments presenting toddlers with symmetric (left) or asymmetric (right) arm behavior. The stickman colorcode indicates the mean angle disparity between forearms/upper arms. **Left:** The absolute angle difference between corresponding parts remained beneath the considered symmetry threshold of $45^o$ [3] (the spike between frames $50-70$ is mostly due to the toddler's regular arm swinging). **Right:** Throughout the 92 frames the angle between the forearms was greater than $100^o$, above the $45^o$ threshold.

90 frames long, or about 3s. In each experiment a single segmentation mask, was obtained interactively in the initial frame. In contrast, Esposito et al. [3] compiled 5-minutes sequences at 8 fps that were manually annotated frame-by-frame using EWMN. Our sequences are shorter, though still sufficient, because our dataset does not contain unsupported gait for longer periods, in part due to the fact that the sessions took place in a small cluttered room (left image in Fig. 5).

From our experiments we have noted that it is often insightful enough to analyze the arm symmetry for the 2D case. Since we are interested in providing measurements for the clinician, Fig. 6 depicts the absolute 2D angle difference between corresponding arm parts (left and right forearms/upper arms) across time in the video segments. From these measurements, different data can be extracted and interpreted by the specialists. In [3] for instance, they look at two different types of symmetry: Static Symmetry (SS) and Dynamic Symmetry (DS). The former assesses each frame individually, while the latter evaluates groups of frames in a half-second window. If at least one frame is asymmetric in a window, then the entire half-second is considered asymmetric for DS. Asymmetry occurs if the angle between two corresponding arm parts differ by more than $45^o$. In our examples the children presented either an entirely symmetric (Fig. 6, left) or asymmetric (Fig. 6, right) behavior, according to both SS and DS.

As in the visual attention tests, both the automatic and expert assessments agree for the studied video segments, although diagnosis takes into account the full session. In the presented segments, the participants were 15 months old. While part. 2 has been diagnosed with autism, part. 1's case was deemed inconclusive and will be followed until 36 months of age.

## IV. Conclusion

This work is the first achieved milestone in a long-term project for early observation of children in order to aid in diagnosis of neurodevelopmental disorders. With the goal of aiding and augmenting the visual analysis capabilities in evaluation and developmental monitoring of ASD, we proposed semi-automatic computer vision tools to observe specific behaviors related to ASD elicited during AOSI, providing both new challenges and opportunities in video analysis. The proposed tools significantly reduce the effort to only requiring interactive initialization in a single frame. We focused on four activities performed during the battery of assessments of development and behaviors related to ASD: three activities were performed during the AOSI and were related to visual attention and one which involves motor patterns observed at any point during the assessment battery. We developed specific algorithms for these activities, obtaining a clinically satisfactory result.

For the visual attention tests, we plan on complementing the estimation of the child's motions with estimating the clinician's movements in order to correlate both. For the assessment of the motor patterns, we will incorporate full 3D information using a richer 3D human model. Of course, there are additional behavioral red flags of ASD, both included in and beyond the scope of AOSI, which we aim at addressing in the future. This includes detecting ASD risk in ordinary classroom and home environments.

## References

[1] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, "Behavioral manifestations of autism in the first year of life," *Int. J. Dev. Neurosci.*, vol. 23, no. 2-3, pp. 143–152, 2005.

[2] R. Landry and S. Bryson, "Impaired disengagement of attention in young children with autism." *J. Child Psychol. Psychiatry*, vol. 45, no. 6, pp. 1115–22, 2004.

[3] G. Esposito, P. Venuti, F. Apicella, and F. Muratori, "Analysis of unsupported gait in toddlers with autism," *Brain Dev.*, vol. 33, no. 5, pp. 367–373, 2011.

[4] P. T. Shattuck, M. Durkin, M. Maenner, C. Newschaffer, D. S. Mandell, L. Wiggins, L.-C. C. Lee, C. Rice, E. Giarelli, R. Kirby, J. Baio, J. Pinto-Martin, and C. Cuniff, "Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study." *JAACAP*, vol. 48, no. 5, pp. 474–483, May 2009.

[5] G. Dawson, "Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder," *Dev. Psychopathol.*, vol. 20, no. 03, pp. 775–803, 2008.

[6] W. Jones, K. Carr, and A. Klin, "Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder," *Arch. Gen. Psyschiatry*, vol. 65, pp. 946–954, 2008.

[7] S. Bryson, L. Zwaigenbaum, J. Brian, W. Roberts, P. Szatmari, V. Rombough, and C. McDermott, "A prospective case series of high-risk infants who developed autism," *J. Autism Dev. Disord.*, vol. 37, pp. 12–24, 2007.

[8] M. Marin-Jimenez, A. Zisserman, and V. Ferrari, ""Here's looking at you, kid." Detecting people looking at each other in videos," in *BMVC*, Dundee, UK, 2011.

[9] B. Lisa and T. Ying-Li, "Comparative study of coarse head pose estimation," in *WMVC*, Orlando, USA, 2002.

[10] J.-M. Odobez and O. Lanz, "Sampling techniques for audio-visual tracking and head pose estimation," in *Multimodal Signal Processing: Human Interactions in Meeting*. Cambridge Univ. Press, 2012.

[11] D. Wen, C. Fang, X. Ding, and T. Zhang, "Development of recognition engine for baby faces," in *ICPR*, Istanbul, Turkey, October 2010.

[12] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-TLD: Tracking-learning-detection applied to faces," in *ICIP*, Hong Kong, China, 2010.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, San Diego, USA, June 2005.

[14] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – Automatic naming of characters in TV video," in *BMVC*, Edinburgh, UK, 2006.
[15] O. Teitelbaum, T. Benton, P. K. Shah, A. Prince, J. L. Kelly, and P. Teitelbaum, "Eshkol-Wachman movement notation in diagnosis: The early detection of Asperger's syndrome," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 32, pp. 11 909–11 914, 2004.
[16] P. Kohli, J. Rihan, M. Bray, and P. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *IJCV*, vol. 79, pp. 285–298, 2008.
[17] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *IJCV*, vol. 99, pp. 190–214, 2012.
[18] P. A. V. Miranda, A. X. Falcão, and J. K. Udupa, "Cloud models: Their construction and employment in automatic MRI segmentation of the brain," IC, University of Campinas, Tech. Rep. IC-10-08, March 2010.
[19] A. X. Falcão, J. Stolfi, and R. A. Lotufo, "The image foresting transform: theory, algorithms, and applications," *PAMI*, vol. 26(1), pp. 19–29, 2004.
[20] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video SnapCut: robust video object cutout using localized classifiers," *ACM Trans. Graph.*, vol. 28, pp. 70:1–70:11, July 2009.