


Social Recognition of Joint Attention Cycles in Children With Autism Spectrum Disorders

Jingjing Liu , Zhiyong Wang , Haibo Qin, Yi Wang, Jingxin Deng, Huiping Li, Qiong Xu, Xiu Xu, and Honghai Liu , *Fellow, IEEE*

Abstract—Objective: Autism Spectrum Disorders (ASD) are characterized by impairments in joint attention (JA) comprising two components: responding to JA (RJA) and initiating JA (IJA). RJA and IJA are considered two inter-related aspects of JA, related to different stages of infant development. While recent technologies have been used to characterize RJA emerging in earlier childhood, only a limited number of studies have attempted to explore IJA, which progressively becomes evident as a hallmark of ASD. This study aims to achieve the social recognition of both RJA and IJA by vision-based human behavior perception through a multi-modal framework automatically and comprehensively. Methods: The first three layers of this framework leverage localization, feature extraction, and activity recognition. On this basis, three critical activities in JA are recognized: attention estimation, spontaneous pointing, and showing actions. Then different behaviors are linked through the fourth layer, semantic interpretation, to model the JA event. The proposed framework is evaluated on experiments of four groups: 7 children with ASD, 5 children with mental retardation (MR), 5 children with developmental language disorder (DLD), and 3 typically developed

children (TD). Results: Experimental results compared with human codings demonstrate recognition reliability with an intra-class coefficient of 0.959. In addition, statistical analysis suggests significant group difference and correlations. Conclusions: The multi-modal human behavior perception-based framework is a feasible solution for the recognition of joint attention in unconstrained environments. Significance: Thus the proposed approach has the potential to improve the clinical diagnosis of autism by offering quantitative monitoring and statistical analysis.

Index Terms—Gaze estimation, joint attention, movement detection, social behavior disorder.

I. INTRODUCTION

AUTISM, spectrum disorder is a pervasive developmental disorder that has attracted widespread attention due to its high prevalence rate (1/59 [1]). ASD symptoms imply social difficulties (e.g., deficits in social communication and social interaction) and stereotyped behaviors and interests. Among the symptoms of ASD, joint attention is an important behavioral risk marker that describes the ability to coordinate attention with others in social interactions by referring to objects and events in the surrounding environment. The recognition of JA can assist in the early screening of autism or improve the social skills of children during the intervention. Specifically, JA can be mainly broken into two components: responding to joint attention and initiating joint attention. For RJA, the individual responds to another person's gaze shift or gesture to reach the same focus of attention. As for IJA, the individual initiates episodes of shared attention by eye gazing, pointing, or showing. Researches [2], [3] suggest that responding abilities precede the initiation skills during childhood but both of them are significant markers for the typical development of social skills.

Recently, various computational approaches are used to provide quantitative assessments of JA which are more precise and objective than human codings. Among these methods, eye-tracking technology is the most widely used method to analyze visual attention as a key factor of JA. In early studies like [4], eye-tracking results suggested that there were significant group differences at a microstructure level (duration of first fixation) between children with ASD and those with TD. Later the Electroencephalogram (EEG) was introduced in [5] to design a novel brain-computer interface for the classification of social joint attention in autism. Compared with eye-tracking and EEG-based methods, computer vision based methods are more convenient

Manuscript received 15 January 2023; revised 30 April 2023 and 16 June 2023; accepted 10 July 2023. Date of publication 18 July 2023; date of current version 25 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61733011, in part by the Shanghai Key Clinical Disciplines Project and Guangdong Science and Technology Research Council under Grant 2020B1515120064, in part by the National Key Research and Development Program of China under Grant 2022YFC3601700, in part by the National Natural Science Foundation of China under Grant 52275013, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515120064, in part by the Shenzhen Science and Technology Program under Grant JCYJ20210324120214040, and in part by the International Cooperation and Exchange of the National Natural Science Foundation of China under Grant 62261160652. (Corresponding authors: Honghai Liu; Xiu Xu; Qiong Xu.)

Jingjing Liu is with the State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, China, and also with the State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, China.

Zhiyong Wang is with the State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, China.

Haibo Qin is with the State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, China.

Yi Wang, Jingxin Deng, and Huiping Li are with the National Children's Medical Center, Department of Child Health Care, Children's hospital of Fudan University, China.

Qiong Xu and Xiu Xu are with the National Children's Medical Center, Department of Child Health Care, Children's hospital of Fudan University, Shanghai 201102, China (e-mail: xuqiong@fudan.edu.cn; xuxiu@fudan.edu.cn).

Honghai Liu is with the State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Shenzhen 518057, China (e-mail: honghai.liu@icloud.com).

Digital Object Identifier 10.1109/TBME.2023.3296489

and intuitive. The recent advances in vision-based head pose estimation and gaze estimation areas [6], [7], [8] show a great potential to characterize joint attention behavior in an unconstrained environment. Computer vision based methods can also provide more interpretation of human behaviors except for gaze performance. For example, besides the gaze item, the body movements were also measured to explore 4-dimensional (spatial 3D + time) behavior of children's JA performance in [9]. More recently, a camera array-based algorithm was employed to detect the participant's head pose in [10] during the JA intervention of robots. They proposed algorithms to use the head pose to approximate participants' gaze directions based on multiple vision sensors and verified the feasibility. Despite of these advances, more accurate vision-based head pose estimation and gaze estimation methods remain to be applied to model the JA event. Additionally, except for visual attention, hand gestures and movements also need to be measured as significant factors in JA.

Despite of the promising progress of JA analysis using computational approaches, most emerging researches only focus on RJA while only a limited number of works explore the IJA task. Aware of this gap, L Billeci et.al [11] monitored the eye gaze of TD children and children with ASD in both RJA and IJA tasks with the aid of eye tracking devices. Results suggested no differences in RJA for two groups, whereas different patterns of gaze fixation and transition in IJA between the groups. They also took both RJA and IJA into consideration to investigate the correlation between JA and neural circuitries in [12]. Despite of these attempts on both IJA and RJA task, we notice that there is few study that realizes the automatic distinction of both RJA and IJA using vision sensors.

Since joint attention impairment is one of the significant early signs of ASD, most studies take TD children and children with ASD as subjects to analyze their differences in terms of JA patterns. However, it is concluded in early researches [13] that the JA impairments have certain correlations with language development. Furthermore, it is reported that autism is often confused with mental retardation and language delay [14]. However, there are few recent studies exploring the JA performance of children with language delay as well as mental retardation to demonstrate the group differences. In order to enhance the distinction of ASD, incorporating the investigation of MR and DLD groups on JA tasks is also in need.

Regarding existing JA-related works, there are three major limitations: limited target groups only focusing on TD and ASD, insufficient analysis of IJA, and lack of comprehensive human behavior analysis. Aware of these limitations, in this work we focus on the modeling of both the RJA and IJA events in the natural interaction tasks. Except for gaze attention, children's hand gestures and upper body movements are also measured to describe the JA event more comprehensively via a multi-modal framework. Also the presented study incorporates the MR and DLD groups for further analysis.

In this article, we propose a hierarchical framework using data of multiple modalities to model the JA event. In detail, the JA event model incorporates four processing layers: localization, feature extraction, activity recognition and semantic representation. We build a multi-vision sensing system to capture the interactions between children and clinicians from

different angles. Locations of different human body parts are detected at first given the multi-sensory data. Then corresponding data are used for extracting features in terms of gaze, hand gesture and upper body movement respectively. On the basis of these extracted features, key activities in the JA events are recognized by processing with features on successive frames. After obtaining comprehensive human behavior perception, the framework entails the recognition of JA events by knowledge graph technologies. The feasibility of the proposed method is validated by experiments under unconstrained situations. Furthermore, we investigate the JA performance on four groups: ASD, MR, DLD, and TD group. The main contributions of the article can be summarized as follows:

1. A comprehensive study for identification of both responding JA and initiating JA is attempted through a unified framework.
2. A multi-modal framework for perception of ASD children's behavior is presented which has four processing layers: localization, feature extraction, activity recognition and high-level semantic manifestation. Specifically, this is the first study in which a few-shot learning based method within a trajectory-aligned HOF (histograms of optical flow) descriptor is proposed to realize the detection of children's showing actions as an important part of this framework.
3. Quantified JA performance are investigated involving children with ASD, MR, DLD and typically developed children, providing potential interpretability for the development of JA skills.

The rest of this article is organized as follows. Section II presents our multi-modal framework for realizing automatic identification of JA. Section III describes experiments on four groups of children as well as the subsequent results. Section IV presents the discussion about the identification results and statistical analysis. And the article is concluded in Section V with future works.

II. MATERIALS AND METHODS

A. Apparatus

In order to evaluate both the IJA task and RJA task in an unconstrained environment, a multi-vision sensing platform is elaborately designed as shown in Fig. 1. Three RGB sensors (Logitech BRIO C1000E) in conjunction with one RGBD sensor (Microsoft Azure Kinect DK) are used to perceive the interactions between the clinician and the child on the tabletop. These sensors are located at specific positions to ensure the coverage of captured scenes and the unified analysis of data. Except for the sensors, a cartoon poster is hung above the camera C3 as an object to reach joint attention. Also some toys will be provided to the child for play as well as the intermediary for joint attention.

B. Paradigm

Both the IJA task and RJA task constitute the whole paradigm which is set to 5 minutes for each child. In the phase of interactions between the clinician and the child, two trials of RJA task will be initiated by the clinician. For the RJA trial, the clinician calls the child's name continuously at the beginning to

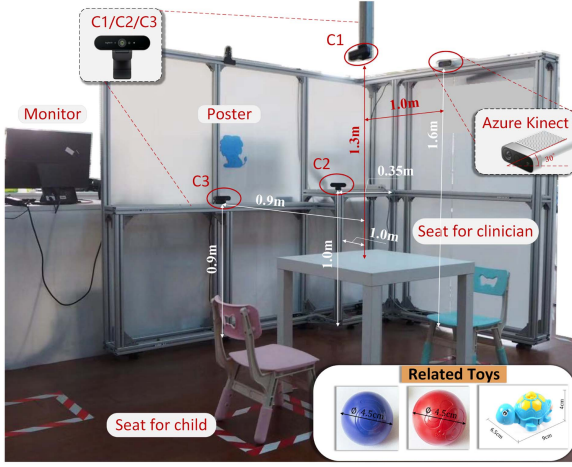


Fig. 1. Multi-sensing platform with marked dimensions. Toys provided to children are also annotated. Camera C1, C2 and C3 are located at the top, left front and flank of the table top respectively. And the Kinect takes an angle of depression of 30° .

secure the child's attention. Then the clinician will try to direct the child's attention to the cartoon poster by gaze shift, pointing gesture and speech. The child's behavior will be observed in the following 4 seconds to evaluate whether the child has response to the JA initiated by the clinician. Since the IJA task implies spontaneous behavior, it could occur at any time during the experimental process. It is defined as two patterns: IJA1 and IJA2. The IJA1 task implies the child has spontaneous gaze at the clinician and pointing gestures while the IJA2 task is identified as combination of spontaneous gaze and showing actions.

C. Methods

In all phase the four vision sensors record the interactive scenes synchronously through multi-thread programming. Within the acquired video streams from different sensors, we propose a hierarchical framework to describe the JA event model. The pipeline of the proposed approach is depicted in Fig. 2.

1) Localization: Given the images captured by the multi-sensing platform, a preliminary disposal as well as the localization of some critical positions of the human body is implemented.

For facial landmark detection, a robust facial landmark detection method [15] is employed for the video stream from the Kinect or camera C2 to detect the child's face.

As for object detection, we used the YOLOv3 model [16] to detect the locations of four objects: two kinds of toys, the child's hands and the clinician's hands. The model pretrained on the VOC2007 dataset [17] is trained on our own dataset including more than 15000 images performed by ten subjects. The network is trained for 100 epochs using a RMSprop Optimizer [18] with the learning rate of 1×10^{-3} in the first 50 epochs and 1×10^{-4} in the last 50 epochs.

And for human pose estimation, the Openpose model [19] is applied to the images from C3 and images from C2 to acquire the child's skeleton joints and the clinician's respectively.

2) Feature Extraction: Subsequent features in terms of gaze direction, hand gesture and upper body movement are extracted on the basis of located human body parts.

• Gaze estimation

Eye gaze is typically the most significant feature used in the JA course. Considering the unconstrained environment in our settings, both 3D head pose estimation method and 3D eye gaze estimation method are utilized to decide the final gaze directions of the children as shown in Fig. 2(b). Given the images captured by camera C2, a robust head pose estimator [20] is employed to obtain the child's head poses due to its high accuracy even under extreme head poses, where $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ is the rotation matrix characterized by head pose angles (yaw, pitch, roll). 3D eye gaze estimation is then applied on those images presenting front faces of children which are filtered by the preceding head pose estimation results. The HRNet [21] is leveraged on a subset of gaze estimation dataset ETH-XGaze [22]. We train the network for 25 epochs, using the ADAM optimizer [23]. The initial learning rate is set to 0.0001 and is decayed by a factor of 0.1 every 10 epoches. The trained network combined with the facial landmark detection results are implemented as the predictor of children's eye gaze. To be more specific, the final gaze vector \mathbf{G} in the world coordinate system as well as camera C2's coordinate system is determined as:

if ($pitch < 10^\circ$) :

$$\mathbf{G} = \mathbf{R}(K \rightarrow C2) \cdot \mathbf{G}_K$$

elseif ($yaw < 45^\circ$) :

$$\mathbf{G} = \mathbf{G}_{C2}$$

else :

$$\mathbf{G} = \mathbf{R} \cdot [0 \ 0 \ -1]^T \quad (1)$$

where $pitch, yaw$ denote the detected head pose angles using the images captured by camera C2, and $\mathbf{R}(K \rightarrow C2)$ is the rotation matrix from the coordinate system of Kinect to the camera C2. If $pitch < 10^\circ$, the gaze direction is calculated using the images captured by the Kinect as \mathbf{G}_K and then \mathbf{G}_K is transformed into the coordinate system of camera C2. If $yaw < 45^\circ$, the gaze direction is calculated using the images captured by camera C2. Otherwise, the head pose orientation is taken as the approximate estimation of the gaze direction.

• Hand gesture recognition

After obtaining the cropped images of human hands, a binary classification Resnet model [24] is employed to detect the pointing gesture which is critical in the JA tasks. Cropped images of both the clinician's hands and child's hands are used in one single recognition model. Certain scaling operations are administered on the images of children's hands to reach the same scale as the clinician's hands. However, due to the low frequency of using the pointing gesture, we employ 10 performers and they are required to perform more forms of pointing gesture beyond

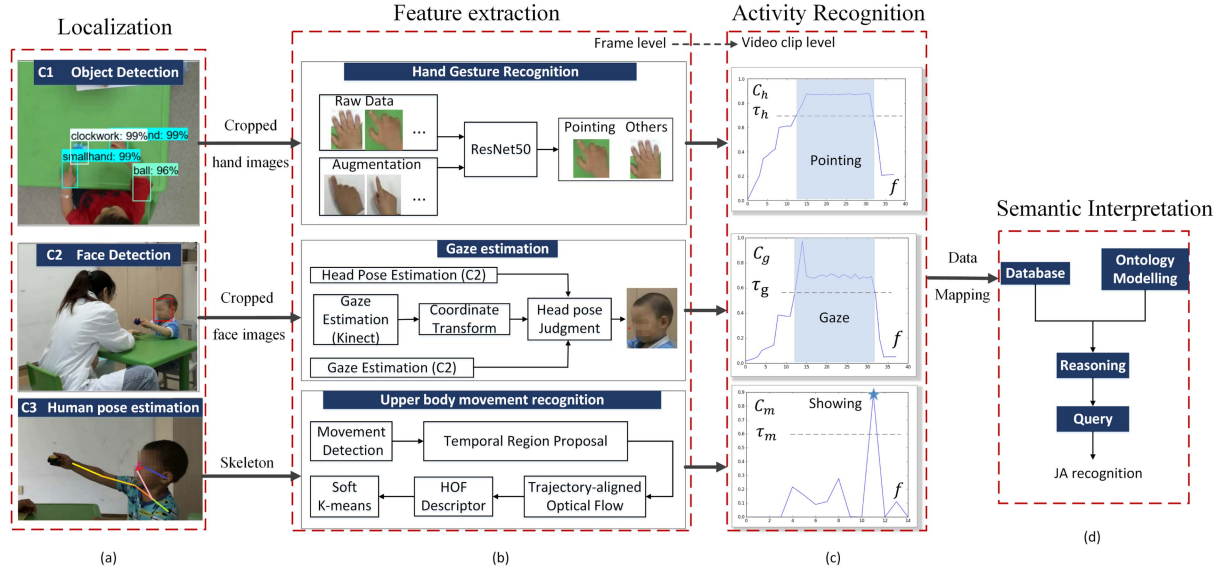


Fig. 2. Overview of the pipeline for JA event detection. A hierarchical model is presented which incorporates four layers. Firstly, the *Localization* layer aims to localize the pivotal regions of human body parts (ROI) in the images from different cameras. Secondly, given the located human body parts, the *Feature Extraction* layer focuses on calculating the relevant features in the ROI to describe the characteristics of the human behavior for each frame. Thirdly, the *Activity Recognition* layer provides with the recognition of three key activities in the video stream on the basis of extracted features from each single frame. Finally, a *Semantic Interpretation* layer presents a semantic and interpretable manner to link the activities and actors in the interactive scenes.

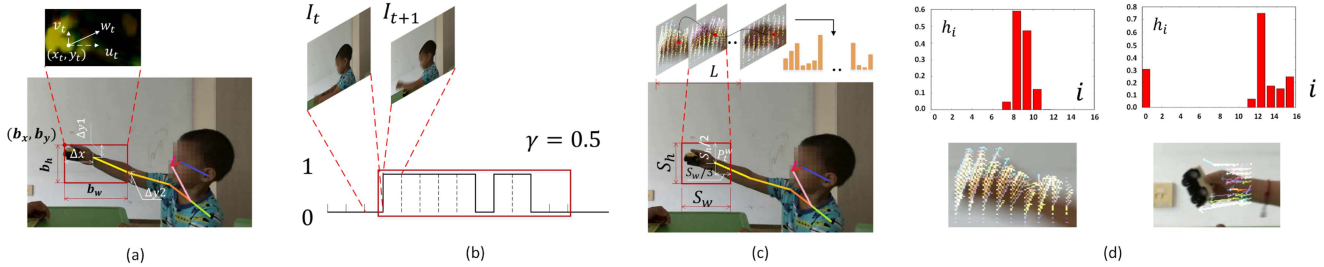


Fig. 3. (a) Dense optical flow calculation in the ROI to distinguish the frames within significant movement. The bounding box of the ROI is determined as: $b_w = \|P_{t,x}^w - P_{t,x}^e\| + \Delta x$, $b_h = \|P_{t,y}^w - P_{t,y}^e\| + \Delta y1 + \Delta y2$, $b_x = P_{t,x}^e - b_w$, $b_y = \min(P_{t,x}^w, P_{t,x}^e) - \Delta y1$, where $(P_{t,x}^w, P_{t,y}^w)$ and $(P_{t,x}^e, P_{t,y}^e)$ denote the positions of the child's wrist and elbow joints respectively. (b) An illustration of the grouping of the temporal region proposal when γ is set to 0.5. The red rounding box denotes one generated region proposal. (c) Trajectory-aligned HOF calculation in the neighborhood of child's wrist joints. (d) The histograms corresponding to two cluster centers of the showing action are presented. And for each cluster, the dense optical flow of one sample is depicted to demonstrate the motion trend.

the experiment. The corresponding images as well as additional operations such as flip and rotation are also adopted for data augmentation. As a result our hand gesture dataset consists of more than 6000 images. The pretrained Resnet model on ImageNet dataset [25] is fine tuned on our own dataset for 50 epochs with a learning rate of 0.1, a weight decay of 5×10^{-4} and a momentum of 0.9.

- Upper body movement detection

In previous works, only gaze related items were considered to assess the IJA task. In this study, we also aim to detect the showing actions of children as they occur in the IJA tasks. While such samples account for a small fraction of possible actions in the interactions, we creatively propose a specific and effective approach to detect the showing action as shown in Fig. 2(b). To our knowledge, this is the

first study which proposes a method to detect the children's showing actions.

Raw video stream from camera C3 is employed for the movement detection because that the showing action can be easier to distinguish from most other actions in profile. Specifically, it begins with identifying the significant movement performed. Given the human pose estimation results acquired by the previous section, the child's forearm is defined as a bounding box (i.e., b_x, b_y, b_w and b_h) according to the locations of relevant skeleton joints as depicted in Fig. 3(a). Then the dense optical flow field $w_t = (u_t, v_t)$ is calculated for the ROI (region of interest) in each pair of successive frames I_t and I_{t+1} , where u_t and v_t are the horizontal and vertical components of the optical flow. The frame I_t is labeled as "1" for significant

movement if only the amplitude and range of motion are above certain thresholds as illustrated in (2). Or else, the frame I_t is labeled as “0” for background.

$$\max(w_t | (x_t, y_t)) \geq \tau_w, \\ \left(\sum \text{sgn}(w_t | (x_t, y_t) - \tau_w/10) \right) \geq \tau_r \quad (2)$$

where (x_t, y_t) denotes a point in the ROI.

Given the frame label, a temporal region proposal method [26] is referred to generate action proposals as shown in Fig. 3(b). The final collected proposal set will be pruned using non-maximal suppression with an IoU (Intersection over Union) threshold 0.95.

With a set of candidate temporal regions, the next stage is to distinguish the showing action instances from other kind of actions. This is accomplished by a few-shot learning method within a trajectory-aligned HOF descriptor. Only the optical flow field around the child’s wrist joints P_t^w is considering for detecting actions. And only the estimated joints whose confidence score is above a certain threshold s are allowed for next steps. As for those imprecise estimation results, they are replaced by tracking previous locations using a median filter. Given the trajectory of the wrist joints, we compute a HOF descriptor within a space-time volume aligned with the trajectory to embed the motion information. Empirically, we set the temporal scale to $L = 5$ frames and the size of spatial neighborhood to $S_w \times S_h$ pixels. For the calculated dense optical flow fields of L frames, their magnitudes are used to vote for one of n bins depending on the corresponding orientations ($n = 16$ in our method).

Considering the scarce showing action data, a simple but effective detection way based on soft k-means is developed. Specifically, the HOF descriptors for all showing action data are divided into: 1) the training set $D_1 = \{X_1, X_2, \dots, X_k\}$ containing k samples and 2) the test set $D_2 = \{X_{k+1}, X_{k+2}, \dots, X_{k+m}\}$ containing the left m examples, where $X_i = (x_i^1, x_i^2, \dots, x_i^n)$ denotes the $n - \text{dim}$ descriptor. We conduct a cluster analysis on the set D_1 using k-means clustering algorithm with different settings of class number. We employ the silhouette coefficient to determine the number of clusters used for the k-means clustering algorithm. Specifically, we randomly select partial data as the sub-dataset and then calculate the silhouette coefficient values when the cluster number is set to 2, 3 and 4. This process is repeated for six times. It is suggested that when the number of clusters is set to 2, the average silhouette coefficient is largest corresponding to a better clustering effect. As shown in Fig. 3(d), the set D_1 can be well grouped into two clusters suggesting two kinds of common actions to show an object: one is holding the object high and another is forward handing over the object. Then, a similarity metric is proposed to evaluate the distance from an instance X_j to the cluster

center $F_i = (f_i^1, f_i^2, \dots, f_i^n)$, $i = 1, 2$.

$$e(X_j, F_i) = \sum_{p=1}^n \sum_{q=1}^n x_p \cdot f_q \cdot e^{-\text{dist}(p,q,n)}$$

$$\text{dist}(p, q, n) = \min(|p - q|, n - |p - q|) \quad (3)$$

According to the similarity metric, an instance X_j belongs to the cluster centered at F_i if the similarity $e(X_j, F_i)$ reaches a certain threshold. The calculated HOF descriptor subsequently is taken to compute the distance from the cluster center F_1 and F_2 respectively.

3) Activity Recognition: As the features are extracted at a frame level, we assign corresponding confidence scores to them to recognize the relevant activity at a video clip level. In the course of JA, three kinds of critical activities are included: attention estimation, spontaneous pointing gestures and showing actions. Accordingly, we define a confidence score for each frame to denote the possibility to be recognized as contributing to the key activities as shown in Fig. 2(c).

A 2D gaze direction φ is derived from the given estimated 3D gaze vector $\mathbf{G} = \{x_G, y_G, z_G\}$. The angle φ is modified based on the azimuth angle to a continuous range $[0, 2\pi]$.

$$\varphi = \begin{cases} a \tan 2(y_G, x_G) (y_G \geq 0) \\ a \tan 2(y_G, x_G) + 2\pi (y_G < 0) \end{cases} \quad (4)$$

Then, the 2D gaze angle φ gets a confidence score C_g as shown in (5).

$$C_g = e^{-\frac{|\varphi - a|}{b}} \quad (5)$$

where a and b are radian values defined artificially. They are represented as a_1, b_1 and a_2, b_2 , corresponding to circumstances taking the poster and clinician as the child’s attention target respectively. The parameters a_1, b_1 and b_2 are set empirically. As for the parameters a_2 , which represent that the clinician is taken as the target of child’s attention, they are calculated as follows since the position of clinician is not stable. By applying human pose estimation on images from camera C2, positions of clinician’s right ear (x_r, y_r) and child’s left eye (x_{le}, y_{le}) , right eye (x_{re}, y_{re}) are used to represent to line of sight from the child towards the clinician. Consequently, the parameter a_2 is calculated as $a \tan 2(y_r - (y_{le} + y_{re})/2, x_r - (x_{le} + x_{re})/2)$.

For the recognized hand gesture in each frame, the confidence score C_h is just defined as the confidence score given by the classification model. Similarly, as for the detected body movement, the confidence score C_m is derived from the similarity metric $e(X_j, F_i)$.

$$C_m = \max(e(X, F_1), e(X, F_2)) \quad (6)$$

where X is the HOF descriptor calculated from the video clip started from the current frame with a duration of L frames.

With the sequences of the features’ confidence scores, the activity recognition is conducted by comparing the confidence scores with pre-defined threshold values along the timeline. As shown in Fig. 2(c), when the confidence score C_g, C_h, C_m reach the corresponding thresholds τ_g, τ_h and τ_m respectively, the activity fragments are generated following the temporal region

TABLE I
PARAMETER VALUES

Parameter	Value	Parameter	Value	Parameter	Value
Δx	60	n	16	s	0.1
$\Delta y1 = \Delta y2$	30	a1	$7\pi/8$	S_w	90
τ_w	10	b1	$\pi/6$	S_h	60
τ_r	0.15	b2	$\pi/10$	τ_h	0.5
γ	0.5	τ_g	0.6	τ_m	0.6

and those with known vision or hearing deficits were excluded in advance (approved by Ethics committee of Children's Hospital of Fudan University). The clinical diagnosis were executed by two professional clinicians in advance. During the experimental process, one professional clinician interact with the child according to the designed paradigm. After the experiments, subjective coding results were given by two professional clinicians via watching the recorded videos. Their coding results remained consistent for all the participants in the experiments. The experimental procedures have passed the ethical review.

B. Parameter Setting

As shown in Table I, the details of predefined parameters are listed. Among these parameters, some are defined empirically (in the white cells) while others (in the gray cells) are defined after the preliminary analysis of the group differences. Taking the parameter τ_g as an example, the threshold value is determined after analysis as follows.

After feature extraction applied on each frame, the time series of the confidence score C_g of subject i are obtained as $C_g^i = C_g^{i,1}, C_g^{i,2}, \dots, C_g^{i,n_i}$, where $C_g^{i,j}$ denotes the confidence score of j th frame for subject i . By referring to the threshold selection method in [29], the threshold τ_g is calculated to maximize the separability of the resultant classes. Firstly, assuming a threshold τ_x , subsequently the confidence scores are separated into two groups: C_g^0 and C_g^1 following (7).

$$\begin{aligned} C_g^0 &= \forall C_g^{i,j}, \text{ if } C_g^{i,j} \leq \tau_x \\ C_g^1 &= \forall C_g^{i,j}, \text{ if } C_g^{i,j} > \tau_x \end{aligned} \quad (7)$$

$$\begin{aligned} \mu &= w_0 \cdot \mu_0 + w_1 \cdot \mu_1 \\ w_0 &= N_0 / ((N_0 + N_1)) \\ w_1 &= N_1 / ((N_0 + N_1)) \end{aligned} \quad (8)$$

Then the mean value of all confidence scores μ is calculated as (8). The number of elements in C_g^0 and C_g^1 are N_0 and N_1 respectively. w_0 and w_1 are the proportion of C_g^0 and C_g^1 respectively. And μ_0 and μ_1 are the mean values of C_g^0 and C_g^1 respectively. Next the variance between classes is calculated as (9).

$$g = w_0 \cdot (\mu_0 - \mu)^2 + w_1 \cdot (\mu_1 - \mu)^2 = w_0 \cdot w_1 \cdot (\mu_0 - \mu_1)^2 \quad (9)$$

The core idea of the threshold selection in [29] is to maximize g . Specifically, we traverse a set of values $[0, 0.1, 0.2, 0.3, \dots, 0.9, 1.0]$ with the step of 0.1 for the threshold τ_g , and choose the value corresponding to a maximum value

of the between class variance g as the final threshold. Similar procedures are also applied to τ_h , τ_m , τ_w and τ_r .

C. Results

1) Validation Results: To demonstrate the validity of the proposed algorithms, a preliminary assessment of the feature extraction methods was performed.

- The result of gaze estimation

To test the accuracy of the gaze estimation method in our settings, we employ 8 normal adults to be seated in our environment. The subject was instructed to look at the cross mark on the paperboard with free head poses. The 3D location of the paperboard was changed artificially for 14 sessions and 14 key frames of different head poses were correspondingly selected to test the gaze estimation algorithm. The average angular error of the proposed gaze estimation method was 8.7° .

- The result of hand gesture recognition

Classification accuracy of two kinds of hand gestures in our own hand gesture dataset has reached as high as 98%.

- The result of upper body movement detection

Specifically, the definition for the calculation of Average Precision (AP) in the MEXaction2 dataset [30] is used as the metric for evaluating the results of detecting the showing action. It is reported that the AP of our movement detection method has reached 0.697.

2) Identification Results: For RJA task, the identification results are reported in Table II. Except from the experimental results of the clinician and the proposed method, we also try to compare the results with the vision-based head pose estimation methods in [10] and [31]. They both employed the supervised descent method to compute the head pose of children for characterizing joint attention. As shown in Table III, RJA1 and RJA2 denote two trials of the RJA task. Result sample A/B/C under these items is referred as the result A given by the clinician, the result B given by our method and the result C given by the compared method. The results A, B and C can be either 0 (none-response joint attention) or 1 (having response joint attention) according to the child's performance. As shown in Table IV, the participants #15 and #18 didn't respond to joint attention in the first trial and responded to joint attention in the second trial. It is noted that children's performances are not necessarily consistent between these two trials. The child may not respond to each time of joint attention initiated by others. Compared with the other method, the proposed method is more consistent with the human codings. This is reasonable because the proposed method benefits from the combination of gaze estimation and more accurate head pose estimation results. Taking the human codings as ground truth, assessment by our method fails for several times and the explanations are as follows. For participant #2 and participant #14, the children only glanced at the poster. Such short attention coordination did fulfill the requirements illustrated above, but it wasn't recognized by the clinician due to limited human perception ability. More comprehensive analysis of the obtained gaze directions such as the fixations and saccades should be applied to reveal the visual attention of child.

TABLE II
EXPERIMENTAL RESULTS FOR RJA TASK

Group	ASD							TD			MR					DLD				
Part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RJA1	0/0/0	0/1/1	0/0/0	1/1/1	0/0/1	0/0/0	0/0/0	1/1/1	1/1/1	1/1/1	0/0/0	0/0/0	1/1/1	0/0/1	0/0/0	1/1/0	0/0/0	0/0/0	1/1/1	1/1/1
RJA2	0/0/0	0/1/0	0/0/0	1/1/1	0/0/1	0/0/0	0/0/0	1/1/1	1/1/1	1/1/1	0/0/0	0/0/0	1/1/1	0/1/1	1/1/1	1/1/1	0/0/0	1/1/1	1/1/1	1/1/1

TABLE III
EXPERIMENTAL RESULTS FOR IJA TASK

Group	ASD							TD			MR					DLD				
Part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
IJA1	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0	2/2	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	2/2	0/0
IJA2	0/0	0/0	3/2	0/0	0/0	0/0	1/1	4/4	1/1	3/3	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	1/2	3/3

TABLE IV
RESPONSE LATENCY MEASUREMENTS IN THE RJA TASK

Group	ASD		TD		MR		DLD			
Part.	4	8	9	10	13	15	16	18	19	20
Response latency (s)	1.36	0.88	0.88	0.84	1.68	1.60	1.00	1.36	0.52	0.92

For IJA task, the experimental results are reported in Table III. IJA1 and IJA2 represent two kinds of IJA task as illustrated before. Result sample A/B under is referred as the result A given by the clinician and the result B given by our method. The results A and B indicate the number of times of the appearance of this event. As for IJA task, false negative instances are mainly caused by the children's fast motion. In detail, the pointing gestures of participant of #15 wasn't recognized due to the blurred images in a fast motion. And the calculation of optical flow could be incorrect given fast motions for participant #3 and #19. Consequently, more robust feature extraction methods in terms of hand gesture recognition and movement detection should be prompted.

For all the JA tasks, the inter-rater reliability between clinical diagnosis and our method is excellent which is quantified with two-way intra-class coefficient (ICC) (ICC 0.959, 95% confidence interval 0.929-0.977). In addition, the sensitivity of our method on all the JA tasks achieves 95.1% and the specificity is 92.2%.

3) Statistical Analysis: Statistical analysis concerned some quantitative measurements which are imperceptible by human observations. The first set of analyses investigated the diversities of performances in JA tasks for different groups using Kruskal-Wallis test. In terms of low-level joint attention behaviors as well as RJA, there are significant differences between ASD group and TD group ($P = 0.032$). As shown in Table II, all the typically developing children had responses to the joint attention, demonstrating the normal responding JA abilities of this group. By contrast, most children with ASD failed in the RJA task which verified similar findings in [32]. However, there are not significant differences between other pairs of groups in the RJA task (ASD versus MR, TD versus DLD, MR versus DLD and so on) whose P-value are all larger than 0.05. As for the high-level joint attention behaviors as well as IJA2, it is observed that the differences of TD group and MR group remain significant ($P = 0.021$).

Except for the analysis of significant differences between groups, we tried to delve into response patterns in the RJA

task. As shown in the Table IV, response latency in RJA task was measured as the latency between the start of clinician's spontaneous pointing gesture and the start of child's visual attention shift. In detail, if the temporal regions of the clinician's pointing activity and the child's visual attention on the poster are $[x_1, y_1]$ and $[x_2, y_2]$ respectively, then the response latency is denoted as $x_2 - x_1$. Here we explored the correlations between the response latency and the JA performances using Spearman correlations. Without distinguishing groups, there was a significant negative correlation between the response latency and the IJA times ($r = -0.796$, $P = 0.006$). On the other hand, the duration of the children's visual attention on the clinician is accumulated as the social gaze. And a significant positive correlation ($r = 0.500$, $P = 0.025$) was found between the social gaze accumulation and the JA performance (the sum of RJA times and IJA times).

IV. DISCUSSION

The present study realized the automatic recognition of two components of JA (RJA and IJA) and quantified the performance of children from four groups (ASD, TD, MR, and DLD). There is a high ICC between the results given by our automatic method and the clinical diagnosis which suggests the feasibility of the proposed method. Consequently, the proposed method has the potential to be leveraged in practical applications such as the joint attention intervention robots. By employing an objective and reliable method which can quantify the childrens JA behaviors, the robot could assist in improving the joint attention of children with autism via interactions.

Except for presenting the results of our method, we also want to compare with existing methods. We compared the RJA recognition results using the head pose estimation method in [10], [31]. Taking the human codings as ground truth, our method is more effective than the compared method. Explanations could be given from two aspects. One reason is that the proposed method directly computes the eye gaze when applicable while the compared method approximates the eye gaze using head orientations, so the latter method may fail when the eye gaze directions is inconsistent with head pose in some cases. Another reason is that the proposed method benefits from the pre-trained head pose estimation networks on large-scale public datasets while other JA recognition methods employed those less effective head pose estimation algorithms proposed in the early stage. In detail, the

supervised descent method for solving nonlinear least squares is adopted in both [10] and [31] to compute the head pose. As for IJA, there is rare study using vision-based methods to realize the recognition so it's difficult to make an exact comparison. Specifically, the recognition of IJA mainly concerns the pointing gesture and showing movement detection. However, those deep learning-based gesture recognition methods and movement detection methods with high accuracies on public datasets cannot be directly applied in the JA settings since they require large amounts of data to train the networks. Although our method is traditional based on hand-crafted features due to the limited data size, it provides pioneering insights about the IJA recognition.

In addition, we also conducted the statistical analysis given the quantified performance of JA. The main findings can be summarized as four points. (a) For RJA, there were significant differences between ASD group and TD group. This is consistent with many existing studies. Similarly, a significant main effect of the group on RJA performance was also reported in [33], with TD children performing significantly higher than the children with ASD. In [31], the analysis showed that children with TD responded more than children with ASD to the JA induction performed by the robot, in terms of head movements as responses to JA induction. However, in some early research [34] it was reported that children with DLD responded correctly to joint attention more often than children with ASD. But our results of Kruskal-Wallis test didn't suggest the significant difference between the group of ASD and DLD. This may result from the limited sample size which requires more data of a larger sample size for a deeper analysis. (b) for IJA2, it was observed that the differences of TD group and MR group remained significant; Relevant findings were reported in [35] that groups of higher MA (mental age) displayed more joint attention behaviors than the low MA groups did. (c) there was a significant negative correlation between the response latency in RJA and the IJA times; This gave us hypothesis that there were consistencies between the RJA performance and IJA performance. And in [36], relevant findings indicated that high-level IJA (showing or pointing) was associated with RJA (gaze following). But further evidences remain to be explored to make clear the exact correlations. (d) a significant positive correlation was found between children's social gaze accumulation and the JA performance. It is reasonable since that gaze at the social stimuli, i.e., the human face, is a prerequisite for joint attention behaviour. Overall, the statistical analysis could provide some objective findings at a statistical level which could help in analyzing the relevant mechanism in clinical analysis. For example, the group difference between MR group and TD group could verify the potential link between MA and JA performance in other clinical studies.

Despite of the promising results and potential applications, there are some limitations to be addressed in our future work. Firstly, since the study was designed to realize the social recognition of joint attention using a novel framework, only a small number of patients were recruited and thus the main limitation is the small sample size. In the future, we will enlarge the sample size to enhance the generalizability of the proposed method. Secondly, the proposed method needs to be improved in terms of two aspects. On the one hand, the accuracies of the

feature extraction methods could be improved to generate more accurate features. On the other hands, the recognition of key activities could be more comprehensive by considering various performances if given more samples. Thirdly, a deeper analysis of JA performance on different disorders is limited in this study. Future studies might consider the longitudinal changes of JA performance by taking the ages of children as a variate.

V. CONCLUSION

In this study, we proposed a hierarchical framework for the modelling of JA events to entail the social recognition of both IJA and RJA in autism. A multi-sensing platform is designed to capture and record the interaction process naturally under an unconstrained environment. A set of computer-vision based approaches were proposed to detect the key elements in the JA circles and a knowledge-driven layer is followed to realize the inference of JA recognition. The validity of proposed architecture were proved through experiments of four groups of children: children with ASD, MR, DLD and TD. Additionally, a preliminary statistical analysis is conducted to reveal group differences and potential correlations. Compared with existing related works about JA, this work incorporates both RJA and IJA through a more comprehensive human behavior analysis with the target subjects extended to four groups. In the future work, the sample size will be enlarged to address current limitations. Given more data of subjects, the technical methods and the statistical analysis will be further investigated. In addition, an interactive user application would also be designed to assist the clinicians with the autistic early screening works.

REFERENCES

- [1] J. Baio et al., "Prevalence of autism spectrum disorder among children aged 8 years - autism developmental disabilities monitoring network, 11 sites, United States, 2014," *MMWR-Morbidity Mortality Weekly Rep.*, vol. 67, no. 6, pp. 1–23, Nov. 2018.
- [2] G. Gredebäck et al., "The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers," *Devlop. Sci.*, vol. 13, no. 6, pp. 839–848, 2010.
- [3] K. Chawarska et al., "Automatic attention cueing through eye movement in 2-year-old children with autism," *Child Develop.*, vol. 74, no. 4, pp. 1108–1122, 2003.
- [4] M. R. Swanson and M. Siller, "Patterns of gaze behavior during an eye-tracking measure of joint attention in typically developing children and children with autism spectrum disorder," *Res. Autism Spectr. Disord.*, vol. 7, no. 9, pp. 1087–1096, 2013.
- [5] C. P. Amaral et al., "A novel brain computer interface for classification of social joint attention in autism and comparison of 3 experimental setups: A feasibility study," *J. Neurosci. Methods*, vol. 290, pp. 105–115, 2017.
- [6] D.-C. Cho and W.-Y. Kim, "Long-range gaze tracking system for large movements," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3432–3440, Dec. 2013.
- [7] N. M. Bakker et al., "Accurate gaze direction measurements with free head movement for strabismus angle estimation," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 11, pp. 3028–3035, Nov. 2013.
- [8] R. U. Haque et al., "Deep convolutional neural networks and transfer learning for measuring cognitive impairment using eye-tracking in a distributed tablet-based environment," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 1, pp. 11–18, Jan. 2021.
- [9] S. M. Anzalone et al., "How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D time) environment during a joint attention induction task with a robot," *Res. Autism Spectr. Disord.*, vol. 8, no. 7, pp. 814–826, 2014.

- [10] Z. Zheng et al., "A randomized controlled trial of an intelligent robotic response to joint attention intervention system," *J. Autism Develop. Disord.*, vol. 50, pp. 2819–2831, 2020.
- [11] L. Billeci et al., "Disentangling the initiation from the response in joint attention: An eye-tracking study in toddlers with autism spectrum disorders," *Transl. Psychiatry*, vol. 6, May 2016, Art. no. e808.
- [12] B. Lucia et al., "An integrated EEG and eye-tracking approach for the study of responding and initiating joint attention in autism spectrum disorders," *Sci. Rep.*, vol. 7, Oct. 2017, Art. no. 13560.
- [13] T. Charman, "Why is joint attention a pivotal skill in autism?," *Philos. Trans. Roy. Soc. London*, vol. 358, no. 1430, pp. 315–324, 2003.
- [14] M. Rutter, "Diagnosis and definition of childhood autism," *J. Autism Childhood Schizophrenia*, vol. 8, no. 2, pp. 139–161, 1978.
- [15] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
- [16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*.
- [17] M. Everingham and J. Winn, "The pascal visual object classes challenge 2007 (VOC2007) development kit," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2006.
- [18] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp: Dividethe gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, pp. 26–31, 2012.
- [19] Z. Cao et al., "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
- [20] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2155–215509.
- [21] K. Sun et al., "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.
- [22] X. Zhang et al., "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 365–381.
- [23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Comput. Sci.*, 2014.
- [24] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [25] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [26] Y. Xiong et al., "A pursuit of temporal accuracy in general activity detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017.
- [27] W. R. V. Hage et al., "Design and use of the simple event model (SEM)," *J. Web Semantics*, vol. 9, no. 2, pp. 128–136, 2011.
- [28] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF," Jan. 2007.
- [29] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
- [30] *Mexaction2 dataset*. 2015. [Online]. Available: <http://mexculture.cnam.fr/xwiki/bin/view/Datasets/Mexactiondataset>
- [31] S. M. Anzalone et al., "Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment," *Pattern Recognit. Lett.*, vol. 118, pp. 42–50, 2019.
- [32] G. Dawson et al., "Early social attention impairments in autism: Social orienting, joint attention, and attention to distress," *Develop. Psychol.*, vol. 40, no. 2, pp. 271–283, 2004.
- [33] H.-L. Cao et al., "Robot-assisted joint attention: A comparative study between children with autism spectrum disorder and typically developing children in interaction with NAO," *IEEE Access*, vol. 8, pp. 223325–223334, 2020.
- [34] K. A. Loveland and S. H. Landry, "Joint attention and language in autism and developmental language delay," *J. Autism Develop. Disord.*, vol. 16, no. 3, pp. 335–349, 1986.
- [35] P. Mundy et al., "Joint attention, developmental level, and symptom presentation in autism," *Develop. Psychopathol.*, vol. 6, no. 03, pp. 389–401, 1994.
- [36] K. E. Pickard and B. R. Ingersoll, "Brief report: High and low level initiations of joint attention, and response to joint attention: Differential relationships with language and imitation," *J. Autism Develop. Disord.*, vol. 45, pp. 262–268, 2015.