2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Video Gesture Analysis for Autism Spectrum Disorder Detection

Andrea Zunino*, Pietro Morerio*, Andrea Cavallo†§, Caterina Ansuini†, Jessica Podda†, Francesca Battaglia¶, Edvige Veneselli¶‖, Cristina Becchio†§ and Vittorio Murino*‡

*Pattern Analysis & Computer Vision (PAVIS) †Cognition, Motion and Neuroscience (C'MON)
Istituto Italiano di Tecnologia, Genova, Italy
{name.surname@iit.it}
‡ University of Verona, Department of Computer Science, Verona, Italy
§ University of Turin, Department of Psychology, Torino, Italy
¶ DINOGMI, University of Genoa, Genova, Italy
‖ Child Neuropsychiatric Unit, G. Gaslini Institute, Genova, Italy

***Abstract*—Autism is a behavioral neurological disorder affecting a significant percentage of worldwide population. It especially starts manifesting at very low ages, but it is difficult to early diagnose it since there is not a specific exam or trial that is able to spot it safely. Its detection is in fact mainly dependent from the medical expertise used to assess the patient behavior during direct interviews. This work aims at providing an automatic objective support to the doctor for the assessment of (early) diagnosis of possible autistic subjects by only using video sequences. The underlying idea and rationale come from the psychological and neuroscience studies claiming that the executions of simple motor acts are different between pathological and healthy subjects, and this can be sufficient to discriminate between them. To this end, we devised an experiment in which we recorded, using a standard video camera, patient and healthy children performing the same simple gesture of grasping a bottle. By only processing the video clips depicting the grasping action using a recurrent deep neural network, we are able to discriminate with a good accuracy between the 2 classes of subjects. The designed deep model is also able to provide a sort of attention map in which the zones in the video of major interest are identified in space and time: this "explains" in a certain way which areas the model deems more relevant to the classification purpose, which could also be used by the doctor to make the diagnosis. In the end, this work constitutes a first step towards the development of an automatic computational system devoted to the early diagnosis of autistic subjects, providing the medical expert of a supportive objective method, potentially simple to use in clinical and also more open settings.**

I. Introduction

Autism Spectrum Disorders (ASD) represent a heterogeneous set of neurodevelopmental disorders characterized by deficits in social communication and reciprocal interactions as well as stereotypic behaviors. The prevalence of ASD has been increasing over the past two decades, with a global population prevalence of about 1% worldwide. Converging evidence indicates that early diagnosis followed by intensive intervention is paramount in order to optimize the outcome of the therapy, especially for children. Yet, diagnosis of ASD remains a complex problem. Typically, it relies on specialist medical expertise with diagnostic instrumentation that depends on interpretative coding of child observations, parent interviews, and manual testing.

Recently, advanced computational and engineering methodologies have been employed to meet the needs of psychology and psychiatry applications [1]. In particular, computer vision (CV) and machine learning (ML) methods have demonstrated quite a success in several areas such as affective computing [2], medical imaging [3], and bioinformatics [4]. Actually, there is a clear aid that CV and ML techniques can provide also in the study of ASD. Being ASD a behavioral disease, the visual observation of healthy subjects and patients executing free or prototypical movements, followed by the modeling of the kinematics of the movements, may lead to discriminate the two classes of subjects, supporting the doctor in the diagnosis of this pathology.

This is actually the ultimate usefulness of CV and ML in autism research, that is, providing efficient, data-driven and robust computer-aided diagnostic algorithms, assisting the doctor with an objective quantitative tool for the evaluation of motor signals of Autism.

In this work, we aim at moving a first step in this direction, by proving that it is indeed possible to disambiguate typically developing vs. autistic subjects using an ML approach operating on video sequences of simple executing gestures. To this end, we propose a new video dataset consisting in a set of video clips of reach-to-grasp actions performed by children with ASD and IQ-matched typically developing (TD) children. Children of the two groups were asked to grasp a bottle, in order to perform four different subsequent actions (placing, pouring, passing to pour, and passing to place). Motivated by recent studies in psychology and neuroscience, we attempt to classify whether actions are performed by a TD or an ASD child, by only processing the part of video data recording the *grasping gesture*. Neuroscientific findings [5], [6] and previous literature [7], [8] have proven that, even at the very onset, motor acts are embedding information about the intention with which they are performed. On the other hand, there is

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.

Fig. 1. Sample frames from a video taken from the dataset. The setup is the same for both ASD and TD subjects.

evidence that autistic spectrum disorders manifest in the early prospective control of movements [9]. Grounding on this evidence, in this work we investigate the extent to which ASD can affect the intentionality in the very first stages of gestures.

To the best of our knowledge, we are the first to cast the problem of autism disease diagnosis in a video-based computational form by analyzing and learning how ASD and TD children perform a same simple gesture.

In this respect, several related works are present in the literature: ML methods have been applied to ASD studies with various modalities, such as quotient-based ASD diagnostic tools [10], Magnetic Resonance Imaging data [11], kinematic data [12], eye tracking data [13], [14], and language acoustic data [15]. Such modalities are quite "invasive" or requires specialized personnel, often affecting the behavior of the subjects under study, and limiting the applicability of the computational tools built on top of the acquired data in real (even clinical) settings. Further, from a technical standpoint, most of these studies relies on handcrafted features and simplified linear models that often fail to handle the complexity of human behaviors in less controlled settings. A few works are overcoming these drawbacks taking advantage of recent computational, "deep" ML approaches. For instance, in [13], a novel method for quantitative and objective diagnosis of ASD using eye tracking and deep neural networks is proposed. The analysis here aims at discovering the differences in eye movement patterns between healthy and ASD people looking at natural scene stimuli.

In the present study instead, we are fully detached from invasive sensory devices (e.g., MRI, microphones, eye tracker, Motion Capture landmarks) and rely only on video observations. We are also not constrained to analyze specific handcrafted features coupled to particular (linear) models, but we adopt a computational deep learning approach, which is able to automatically extract relevant features, proved to be enough powerful to classify ASD and TD participants based on objective patterns of behavior from the sole video recordings. In particular, we devise and test a Long-Short Term Memory (LSTM) network and certify that the automatic diagnosis of autistic children is feasible with a good degree of accuracy. Moreover, this model was designed to also provide a sort of attention map, which identifies – in a sense, "explains" – the areas in the video (localized in space and time) the model deems more relevant to the classification purpose. These maps may potentially provide the medical expert of information she/he may use for the intepretation of the diagnosis.

Ultimately, this work contributes in 1) presenting the first video dataset composed by simple gestures performed by ASD and TD children, and 2) proposing an original *ad hoc* deep learning approach aimed at their discrimination by simply processing the video footage, while also providing feedback to the doctor via the estimated attention map. In this framework, all types of subjects are free to operate in a relatively unconstrained setting while performing a simple gestural act: differently from other sensory modalities, this greatly simplifies the applicability of the system to be potentially used in real clinical environments and, in perspective, also in more open settings as a first screening test.

The rest of the paper is organized as follows. Section II presents the dataset and its acquisition protocol. Section III describes in details the LSTM model adopted for our experimental analysis. Section IV provides the numerical results, their discussion and interpretation. Finally, Section V draws conclusions and sketches future work.

II. The Dataset

A. Participants

Twenty children with ASD (18 males) without accompanying intellectual impairment and twenty typically developing children (TD group: 16 males) were recruited from the Child Neuropsychiatry Unit of the IRCCS Giannina Gaslini Hospital and primary schools in Genova. Groups were matched for age (TD mean $\pm$ Std Dev = 9.5 ± 1.5 years.months; ASD = 9.8 ± 1.5 years.months; $t_{38} = -0.665, p = 0.510$), gender, and Full Scale IQ as measured by the Wechsler Scale of Intelligence (WISC IV [16]) (TD mean $\pm$ Std Dev = 102.8 ± 9.4; ASD = 98.5 ± 11.1; $t_{38} = 1.325, p = 0.193$). The research protocol was approved by the local ethics committee (ASL3 Genovese) and was in accordance with the principles of the Helsinki Declaration [17]. Parents provided written informed consent after receiving a detailed description of the study. Children with ASD were diagnosed according to DSM-5 [18] criteria. The Autism Diagnostic Observation Scale (ADOS-2) [19] and Autism Diagnostic Interview-Revised (ADI-R) [20] were administered by two experienced professionals. All children had normal or corrected-to-normal vision and were screened for exclusion criteria (pharmacological treatment, dyslexia, epilepsy, and any other neurological and psychiatric conditions). Both ASD and TD group were assessed for executive functions abilities by means of the Tower of London (TOL) test [21]. This test revealed no significant differences between TD and ASD children (TD mean $\pm$ Std Dev = 29.35 ± 3.54; ASD = 29.35 ± 2.80; $t_{38} = 0, p = 0.999$). All but two of the children (one in the ASD group and one in

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.

the TD group) were right-handed according to the Edinburgh Handedness Inventory [22].

B. Stimuli and procedure

Children were tested individually in a quiet room. They were seated on a height-adjustable chair, with their right elbow and wrist resting on a table (height = 64 cm; length = 100 cm; width = 60 cm). In order to guarantee a repeatable start position, they were asked to maintain their forearm in a pronated position, with their right arm oriented in the parasagittal plane passing through the shoulder, and their right hand in a semi-pronated position. They were asked to keep their thumb and index fingers closed in a pincer grip on a tape-marked point (at about 7 cm from table edge) on the working table. A plastic bottle filled with water (base diameter= 5 cm; height = 18 cm; weight = 225 g) was positioned on the table at a distance of 44 cm from children's midline. Throughout the entire experimental session, the same female experimenter (co-actor), seated at the opposite side of the table, interacted with the children.

Children were instructed to reach towards and grasp an object (a bottle) to place it into a box (grasp-to-place), to pour some water into a glass (grasp-to-pour), or to pass the bottle to a co-actor (grasp-to-pass), who would then either place the bottle into the box (pass-to-place) or pour some water (pass-to-pour).

Depending on condition, one of two target objects was placed on the table: a box (height = 6 cm; diameter = 10 cm) or a glass (height= 10 cm; diameter= 6.5 cm). For grasp-to-place and grasp-to-pour trials, the target object was located 19 cm away from the bottle. For grasp-to-pass trials, the target object was located closer to the co-actors right hand, 43.5 cm away from the bottle. Children performed a series of 12 consecutive grasps for each condition, making a total of 48 movements. On each trial, children were asked to perform the movement at a natural speed after an auditory tone. The order of block presentation was pseudo-randomized across participants. Before each block, there were 2 practice trials to familiarize children with tasks. To avoid fatigue and lack of attention, children were given a 2 minutes pause at the end of each block. Testing required a single session of approximately 30 min per participant. Movements were filmed from a lateral viewpoint using a Vicon VUE video camera (resolution: 1280 x 720 pixels, 100 frames/sec). The video sequences are exactly trimmed at the instant when the hand grasps the bottle, removing the following part. The videos result very short, the shortest one has 41 frames while the average length is 83 frames. We discard the corrupted acquisitions, collecting a final dataset based on 1837 video sequences[1].

III. The Model

We select a model based on the Long Short Term Memory (LSTM) network [23]. LSTM networks have in fact been proven to effectively codify temporal information of video

[1]https://pavis.iit.it/index.php/datasets/autism-spectrum-disorder-detection-dataset

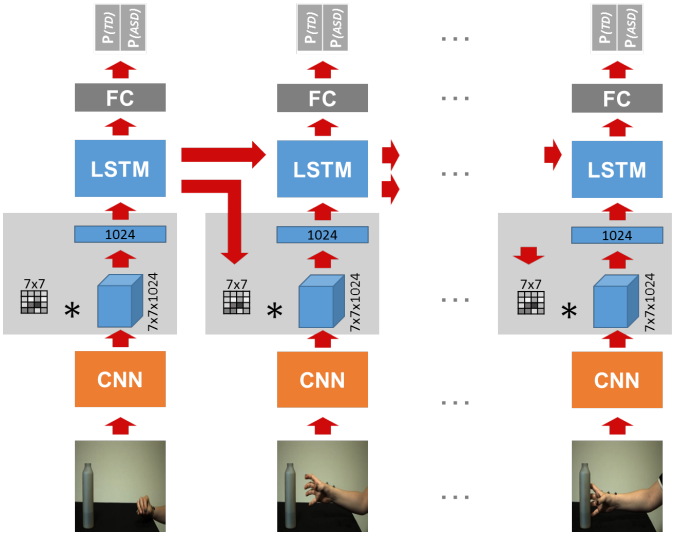

Fig. 2. LSTM-based model used for classifying ASD and TD subjects from video sequences. The top layer (softmax layer) of the model outputs the probabilities $\mathbb{P}(TD)$ and $\mathbb{P}(ASD)$ for the two classes of subject.

Fig. 3. The attention maps superimposed to the video frames. The brighter the white spot, the higher the importance given to those image pixels. (Best viewed on screen).

sequences, especially when fed with powerful deep features [24]. Convolutional Neural Network (CNN) features are first extracted frame by frame from each video and feed recursively the LSTM. We select the features from the last convolutional layer of a pre-trained GoogleNet [25]. Such layer, thanks to its convolutional nature, still preserves some spatial information, thanks to its $7 \times 7 \times 1024$ structure: namely, for each of the 7×7 spatial location, a vector of 1024 feature is extracted by the CNN. This fact is then exploited in an ad-hoc spatial attention mechanism. As depicted in Figure 2, at each time instant t the network weighs the spatial features X_i^t according to a *learned* 7×7 importance map a_i^t, where $i = \{1...K^2\}$, $K = 7$. Finally, a 1024-dim feature vector $x^t = \sum_{i=1}^{K^2} a_i^t X_i^t$ then feeds the LSTM.

The LSTM has a 128-dimensional hidden representation, which is based on the current input and on the past LSTM hidden state. Such 128-dim feature vector is further passed through a non-linear classifier, represented by a 16-dimensional fully connected (FC) layer and a binary softmax. As it can be noted, the 7×7 normalized attention map is also generated by the LSTM, based on its hidden spatio-temporal representation. An internal gating mechanism further allows the LSTM to encode temporal information [26]. As already mentioned, such map is exploited by the net to weigh the $7 \times 7 \times 1024$ feature cuboid extracted at the next frame, in order to produce a single 1024-dimensional feature vector. In

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.

other words, the network learns where to to focus attention at the next time instant. Such soft attention mechanism allows us to inspect (to some spatial extent) *where* the network is learning useful cues within the image frame (Fig. 3). This can provide a further useful visual feedback to clinicians.
As a common procedure for LSTM training, each video is split in 15-frame clips and passed through the entire model which outputs a binary vector containing, for each frame, the probabilities for ASD or TD. During the training of 39 different models[2] each clip is considered independently. The test accuracy for each subject is computed by averaging the scored probabilities over all the frames of the videos.

IV. Experiments

A. Setup

We preliminary process the video frames in order to remove any information of the appearance of the arm, hand that may catch the attention of the model. We apply a Gaussian smoothing over all the images to reduce the details of the visual appearance. In this way we want the algorithm to focus the attention on the kinematics of the gestures only.
We select one-subject-out testing procedure, that is, we compute 39 accuracies, training our system on all the subjects except the one we are testing, then we averaged all the accuracies to get the final classification results.
This testing procedure is more challenging than the usual cross-validation whose classification scores are always higher. We deem the adopted procedure is the correct one to devise a system, effectively able to generalize and detect autistic subjects in real-world applications.

B. Results & discussion

In a first experiment we split the dataset considering all the videos belonging to a single intention at a time. By doing this we consider a lower amount of data, while trying to remove an additional variability factor which could be brought by the different underlying intention. According to this experiment, we obtain four different average (over all the frames of the videos) accuracies for each subject left out. In Table I, we report the accuracies considering a different class of gesture at a time. The first group (first half, subject 1-20) represent the results for ASD subjects and the second group for TD ones (second half, subject 21-40). Each line in Table I refers to a different subject left out. We can clearly note how particular subjects could be perfectly classified as autistic if considering videos belonging to any class (see for example subject 15 in the ASD group). On the contrary, we can also note that a same subject could be almost perfectly classified as autistic if we consider only specific classes of gesture, but not others. For instance, focusing on ASD subject 2, it is evident that this child is easily recognized if he performs the two action pouring and pass-to-pour, while considering the other two classes of action the LSTM model fails. Globally, the obtained accuracy averaged over the intentions is quite high, around 0.69 with a standard deviation (Std Dev in Table I) of 0.22. The standard deviation certifies that a large variability is present between the agents and supports the claim that finding a very general pattern in the video, that could be unique in distinguishing autistic subjects, is a challenging task.
To give a further idea on how difficult the proposed task is, we test dense trajectories [27] which is the hand-crafted state-of-the-art feature in action recognition literature. We adopt the default parameters setup in the code and extract HOG (Histogram of Gradients) [28] and HOF (Histogram of Optical Flow) [29] descriptors. Subsequently, we encode the extracted descriptors with the powerful VLAD encoding [30], as in [8]. As previously done, we conducted the same one-subject-out testing procedure considering one class of gesture at a time. We obtain a random chance accuracy by averaging the four different performances over all the subjects, to demonstrate that dense trajectories encoding is not able to distinguish between gestures performed by ASD and TD subjects. We can state that classical computational methods are not so powerful to finer encode the kinematics of the gestures.
It is known from the literature that deep learning methods require large amount of training data to perform well on unseen samples. In this line, we carried out a second experiment where we considered indistinctly all the videos (actions) belonging to the subjects, no matter which is the underlying intention of the gestures. According to this, the LSTM model has more data to train on, and we could expect a stronger performance in the autism disease classification. The last column in Table I reports the results obtained for each subject left out considering all the data during the training/testing procedure. With this strategy, we can clearly see that the subjects correctly classified are more than the previous case. Precisely, the line Detections ($acc > 0.5$) in Table I (last line at the bottom) highlights which is the number of classified subjects with an accuracy greater than 0.5. The first experiment has correctly classified (as TD or ASD) 28 subjects, while in the second experiment we have improved the performance, obtaining 32 correctly classified children. Consequently, the final detection accuracy in the two cases results 0.72 and 0.82, respectively. The latter score represents a very powerful performance, that clearly certifies that the automatic ASD detection is feasible using a video-based analysis approach only. In Table II, we compute the confusion matrix for the LSTM experiment considering full data. We can note that miss-detected subjects are almost equally distributed between the ASD and TD groups. However, accuracies and standard deviations (Table I) for ASD and TD subjects show some differences: TD subjects are classified with higher accuracy and lower standard deviation (0.77 ± 0.21) with respect to ASD ones (0.72 ± 0.22). This is evidence of the fact that autistic spectrum disorders comprise indeed a very wide spectrum of variability.
To have some insights on the hidden LSTM learning mechanism, following [23] we can overlap the *learned* spatial

[2]Since all the not-corrupted videos belonging to subject 3 refer to the grasp-to-pour condition only, we decide to discard subject 3 in our experiments.

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I

LEAVE-ONE-SUBJECT-OUT VIDEO CLASSIFICATION ACCURACIES OBTAINED WITH LSTM: EXPERIMENTS ARE CARRIED OUT PER CLASS (INTENTION) AND WITH FULL DATA. CLASS 1,2,3,AND 4 REFERS TO THE VIDEOS WITH CONDITIONS PLACING, POURING, PASS-TO-PLACE AND PASS-TO-POUR, RESPECTIVELY. WE HIGHLIGHT IN BOLD THE ACCURACIES GREATER THAN 0.5.

	Subject out	Class 1	Class 2	Class 3	Class 4	Average	Full data
ASD subjects	1	0.67	0.80	0.73	0.27	**0.62**	**0.68**
	2	0.09	1.00	0.08	0.92	**0.52**	**0.74**
	3	-	-	-	-	-	-
	4	0.83	0.64	0.00	0.08	0.39	**0.72**
	5	0.17	0.67	1.00	0.45	**0.57**	**0.68**
	6	0.00	1.00	0.08	0.08	0.29	**0.56**
	7	0.25	0.00	0.83	0.08	0.29	0.15
	8	0.08	0.08	0.42	0.25	0.21	**0.88**
	9	0.92	1.00	1.00	1.00	**0.98**	**0.98**
	10	0.92	1.00	0.58	1.00	**0.88**	**1.00**
	11	1.00	0.92	1.00	1.00	**0.98**	**1.00**
	12	1.00	1.00	0.92	0.42	**0.83**	**1.00**
	13	1.00	0.83	0.83	1.00	**0.92**	**0.73**
	14	1.00	0.75	1.00	0.50	**0.81**	**1.00**
	15	1.00	1.00	1.00	1.00	**1.00**	**1.00**
	16	0.33	0.42	0.17	0.45	0.34	0.43
	17	0.17	0.00	0.09	0.17	0.11	0.16
	18	1.00	1.00	0.83	1.00	**0.96**	**0.89**
	19	0.91	1.00	0.50	1.00	**0.85**	**0.85**
	20	0.75	1.00	1.00	0.92	**0.92**	0.31
Average accuracy ASD		0.64	0.74	0.66	0.61	0.66	0.72
Std Dev ASD		0.35	0.28	0.33	0.35	0.27	0.22
TD subjects	21	0.90	0.75	0.55	0.58	**0.69**	**1.00**
	22	0.33	0.50	0.67	0.08	0.40	0.17
	23	0.42	0.75	0.00	0.00	0.29	0.13
	24	0.58	0.40	0.36	0.42	0.44	0.38
	25	0.64	0.92	0.83	0.50	**0.72**	**0.62**
	26	1.00	1.00	1.00	0.83	**0.96**	**0.91**
	27	0.91	0.75	1.00	1.00	**0.91**	**1.00**
	28	0.67	1.00	1.00	0.92	**0.90**	**0.88**
	29	0.67	0.33	1.00	0.92	**0.73**	**0.89**
	30	1.00	0.75	0.83	0.50	**0.77**	**0.94**
	31	0.92	0.83	1.00	0.92	**0.92**	**0.70**
	32	0.75	0.92	0.91	1.00	**0.89**	**0.94**
	33	0.33	1.00	0.92	0.75	**0.75**	**0.96**
	34	0.50	1.00	1.00	1.00	**0.88**	**0.96**
	35	1.00	1.00	1.00	1.00	**1.00**	**0.94**
	36	0.45	0.25	0.08	0.67	0.36	**0.83**
	37	0.82	1.00	0.82	0.20	**0.71**	**0.68**
	38	0.18	0.50	0.36	0.83	0.47	**0.93**
	39	0.64	0.67	0.45	1.00	**0.69**	**0.61**
	40	1.00	1.00	0.83	1.00	**0.96**	**0.98**
Average accuracy TD		0.69	0.77	0.73	0.71	0.72	0.77
Std Dev TD		0.21	0.20	0.26	0.27	0.17	0.21
Average accuracy		0.66	0.75	0.68	0.66	0.69	0.75
Std Dev		0.30	0.24	0.30	0.30	0.22	0.22
Detections ($acc > 0.5$)		26	30	27	23	28	32
Detection accuracy		0.67	0.77	0.69	0.59	0.72	0.82

TABLE II

CONFUSION MATRIX FOR ASD DETECTION, WITH THRESHOLD 0.5, REFERRED TO THE LSTM MODEL WITH FULL DATA.

	Precision	83.33%	80.95 %	
True	ASD	15	4	78.95%
	TD	3	17	85.00%
		ASD	TD	Recall
		Predicted		

attention map to the video frames and visualize them. In Fig. 3, we superimpose the learned attention map on the frames output from a well-classified test video sample. We can see that the model is coherently classifying the gesture focusing on parts around the hand and the arm of the subject, trying to extract meaningful patterns from the kinematics of this gesture. A supporting clinical instrument represented by this kind of video-based model could not be fully trustable if the scored accuracies are thresholded at 0.5. The LSTM outputs a confidence (probability) of the detection, and we can use this result to perform a sensitivity analysis of the performance. In this respect, we conducted a final investigation where the accuracies obtained with the LSTM model are thresholded in a more restrictive way. Fig. 4 reports LSTM results when the threshold varies from 0.5 to 0.95. We can see that both detection accuracy per class and detection averaged over different class of intentions are quite robust passing from a threshold of 0.5 to 0.7. The two scores decrease from 0.82 to 0.67 and from 0.72 to 0.59, respectively, which still represent powerful performance in automatic autism disease detection. Continuing to increase the threshold, in order to be more confident in the automatic diagnosis, we can state that the model is not more trustable after a threshold of 0.9 where the detection accuracy over full data scores under the random chance. In Fig. 4, we plot the two described detection curves together with the original predicted accuracies which do not vary over the different adopted thresholds.

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.

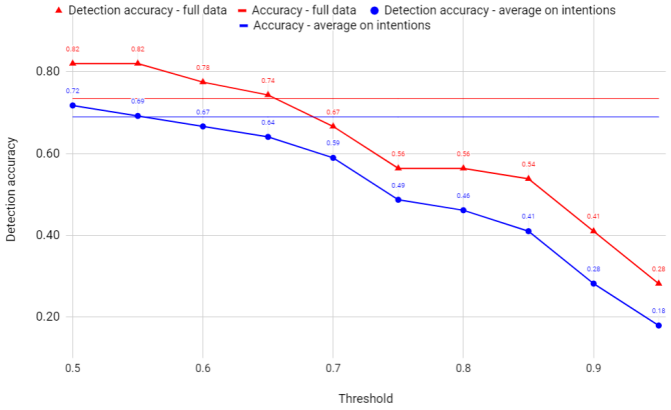

Fig. 4. LSTM detection and average accuracy classification as a function of confidence threshold. The use of a more restrictive threshold results in a decreasing of the detection curves.

V. Conclusions & Future Work

In this paper, we have devised a video-based computational approach to discriminate ASD and TD subjects, which can be regarded as first step towards *objective* autism disease diagnosis. We have proposed a novel dataset of simple grasping gestures labeled according to ASD/TD subject who performed the action and to the different underlying intention. Based on that, we designed an LSTM model with an attentional mechanism, and demonstrated that it is able to classify whether a grasping act is performed by ASD or TD subjects with a robust performance.

As a future work we plan to exploit the additional labels, in order to verify whether intention classification [8] is altered by ASD. The latter fact could be exploited as an additional cue for ASD diagnosis. Moreover we want to move in the direction to improve the results, by refining the spatial attention model to finely inspect the kinematic discriminants in the gestures.

References

[1] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.

[2] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[3] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE transactions on medical imaging*, vol. 24, no. 3, pp. 371–380, 2005.

[4] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.

[5] J. M. Kilner, "More than one pathway to action understanding," *Trends in cognitive sciences*, vol. 15, no. 8, pp. 352–357, 2011.

[6] J. C. Stapel, S. Hunnius, and H. Bekkering, "Online prediction of others actions: the contribution of the target object, action context and movement kinematics," *Psychological research*, vol. 76, no. 4, pp. 434–445, 2012.

[7] A. Zunino, J. Cavazza, A. Koul, A. Cavallo, C. Becchio, and V. Murino, "What will i do next? the intention from motion experiment," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1–8.

[8] ——, "Predicting human intentions from motion cues only: A 2d+ 3d fusion approach," in *ACM on Multimedia Conference*, 2017.

[9] C. Trevarthen and J. T. Delafield-Butt, "Autism as a developmental disorder in intentional movement and affective engagement," *Frontiers in Integrative Neuroscience*, vol. 7, 2013.

[10] P. Kassraian-Fard, C. Matthis, J. H. Balsters, M. H. Maathuis, and N. Wenderoth, "Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example," *Frontiers in Psychiatry*, vol. 7, p. 177, 2016.

[11] C. Ecker, A. Marquand, J. Mourão-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams, and D. G. M. Murphy, "Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," *Journal of Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.

[12] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni, "Use of machine learning to identify children with autism and their motor abnormalities." *Journal of autism and developmental disorders*, vol. 45, no. 7, p. 2146, 2015.

[13] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.

[15] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[16] D. Wechsler, "Wechsler intelligence scale for children." 1949.

[17] G. A. of the World Medical Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects." *The Journal of the American College of Dentists*, vol. 81, no. 3, p. 14, 2014.

[18] D.-. A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders," *Arlington: American Psychiatric Publishing*, 2013.

[19] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, S. Bishop *et al.*, *Autism diagnostic observation schedule: ADOS-2*. Western Psychological Services Los Angeles, CA, 2012.

[20] M. Rutter, A. Le Couteur, C. Lord *et al.*, "Autism diagnostic interview-revised," *Los Angeles, CA: Western Psychological Services*, vol. 29, p. 30, 2003.

[21] P. Anderson, V. Anderson, and G. Lajoie, "The tower of london test: Validation and standardization for pediatric populatons," *The Clinical Neuropsychologist*, vol. 10, no. 1, pp. 54–65, 1996.

[22] R. C. Oldfield, "The assessment and analysis of handedness: the edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, 1971.

[23] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[24] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May 2013.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05, 2005, pp. 886–893.

[29] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[30] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

Authorized licensed use limited to: Amrita School Of Engineering - Kollam. Downloaded on January 28,2024 at 13:20:10 UTC from IEEE Xplore. Restrictions apply.