Video-based Behavior Understanding of Children for Objective Diagnosis of Autism

Abid Ali, Farhood F Negin, Francois F Bremond, Susanne Thümmler

Video-based Behavior Understanding of Children for Objective Diagnosis of Autism

Abid Ali [1,2], Farhood Negin [1] Susanne Thümmler [1,2]
Francois Bremond [1,2]
[1] *Université Cote d'Azur*
[2] *INRIA*
{*f_author, s_author*}*@inria.fr*

Keywords: Autism. Autism-Spectrum-Disorder, action-recognition, computer-vision, 3d-Convolutional-neural-network

Abstract: One of the major diagnostic criteria for Autism Spectrum Disorder (ASD) is the recognition of stereotyped behaviors. However, it primarily relies on parental interviews and clinical observations, which result in a prolonged diagnosis cycle preventing ASD children from timely treatment. To help clinicians speed up the diagnosis process, we propose a computer-vision-based solution. First, we collected and annotated a novel dataset for action recognition tasks in videos of children with ASD in an uncontrolled environment. Second, we propose a multi-modality fusion network based on 3D CNNs. In the first stage of our method, we pre-process the RGB videos to get the ROI (child) using Yolov5 and DeepSORT algorithms. For optical flow extraction, we use the RAFT algorithm. In the second stage, we perform extensive experiments on different deep learning frameworks to propose a baseline. In the last stage, a multi-modality-based late fusion network is proposed to classify and evaluate performance of ASD children. The results revealed that the multi-modality fusion network achieves the best accuracy as compared to other methods. The baseline results also demonstrate the potential of an action-recognition-based system to assist clinicians in a reliable, accurate, and timely diagnosis of ASD disorder.

1 INTRODUCTION

Autism Spectrum Disorder (ASD) represents a heterogeneous set of neurobiological disorders characterized by defects in social communication and reciprocal interactions, repetitive, as well as stereotypic behaviors. ASD onsets in early childhood and have a substantial influence on the lives of children and their families without having a definite treatment. Although researchers associate ASD to various factors such as genetic, biological, and environmental effects, nevertheless, they are unknown in many patients [O'Roak and State, 2008]. Moreover, the prevalence of ASD increases. According to World Health Organization (WHO), 1 out of 160 children has an ASD [WHO, 2021]. This number is the average of various studies where the reported prevalence varies significantly across them. As reported by the Autism and Developmental Disabilities Monitoring (ADDM) in 2016 the prevalence of autism spectrum disorder is now one in 54 children [Knopf, 2020]. Also, the prevalence in middle and low-income countries is unknown.

The common agreement in the literature suggests that early diagnosis, accompanied by continuous intervention, is a key factor to maximize therapeutic results. Therefore, exploiting the brain's neuroplasticity at early childhood, timely diagnosis of ASD and recommending intensive behavioral treatments can result in better long-term outcomes. However, ASD diagnosis still continues to be a complicated challenge. The significant parameters include expert knowledge and particular diagnostic tools based on decoding child behavior, parent interviews, long-term follow-ups and inspection of symptoms, and manual analysis. These assessments are time-consuming and clinically require arduous processes. Additionally, human assessments are subjective and inconsistent. Early studies suggested that abnormalities in social interactions, communication, and presenting repetitive behaviors could be the primary indicatives of ASD [Kanner et al., 1943]. However, some of the autism-related communicative and behavioral symptoms are not exclusive to autism, which makes getting an early diagnosis difficult [Lewis and Bodfish, 1998]. Proper treatment requires timely diagnosis though, the accu-

rate diagnoses were generally made not sooner than age 5, which is already late for treatment [Hashemi et al., 2012].

Computer vision and computational behavior modeling are concerned with machine analysis and understanding human behavior. They help to develop analytic techniques to automate diagnostic assessments. These approaches could facilitate the quick analysis of vast amounts of naturally recorded videos in addition to supporting the diagnosis process by providing accurate and objective measurements. Video-based behavior understanding could enable the improvement of current assessment protocols and the discovery of new behavioral markers. Specifically, action recognition techniques can model the motor patterns of children and spot the presence of typical behavior related to gait, posture and repetitive motions [Marinoiu et al., 2018, Zhang et al., 2021, de Belen et al., 2020, Pandey et al., 2020]. There is no specific test to spot these atypical patterns and clinicians take a holistic approach during their sessions with children to spot those patterns in their behaviors. Therefore, a video-based action recognition system is beneficial in carrying out surveillance of children in the clinical sessions and detect abnormal behavioral patterns. From an action recognition point of view, such analysis introduces a number of challenges that usually do not arise in other action recognition scenarios. For example, repetitive behavior such as stereotypical self-stimulatory behavior which is considered as a significant clue of this disorder does not comprise any rule or constraint that are common in regular action datasets. Moreover, to have a reliable assessment of the subjects, the actions, communication, and interactions need to be captured naturally. However, in clinical settings, children consider clinicians as an authority and do not act in an unaffected manner [Huerta and Lord, 2012].

In order to capture all subtleties associated with the diagnosis of ASD and to explicitly model a child's behavior in a natural setting, collecting standard datasets is indispensable. Currently, there is a severe lack of data in this domain, which hinders rapid progress. Most of the available public datasets for diagnosis of ASD centered on emotional involvement, facial expression and eye movement [Wang et al., 2015b, Tanaka and Sung, 2016] and less focused on motion perception and action recognition aspects [Rajagopalan et al., 2013, Negin et al., 2021]. To study subjects' behavior from videos in an uncontrolled natural environment we have collected a novel dataset from children attending the clinical center and willing to take part in our studies.

Considering the motivation and above-mentioned challenges, we attempt to propose a computer-assisted solution for automatic analysis of Autism Spectrum Disorder (ASD) behavior using action recognition approaches. The contributions are as follows:

- We propose a new framework based on multi-modality fusion for recognizing autistic behaviors in videos. The method is based on I3D architecture pre-trained on a large-scale action recognition dataset and fine-tuned on a small dataset of stereotypic actions. We study both the RGB and optical flow modalities for their contribution in recognizing ASD actions. Later on, we fuse the two modalities to achieve a higher accuracy.
- To evaluate our proposed methodology, we attempt to collect and annotate a novel Autism dataset, **Activis**, in an uncontrolled environment with the help of clinicians.
- We also compare different action recognition networks and produce baseline results for the collected dataset. The multi-modality fusion network achieved higher accuracy as compared to the rest of methods.

In the rest of the paper, the state-of-the-art methods are discussed in Section 1.1. In section 2 we propose our method and the necessary pre-processing steps, followed by details of the dataset and experiments conducted with results in Section 3. Finally, we conclude the paper in Section 4

1.1 Related Work

To diagnose ASD, psychologists designed a standard semi-structured test called Autism Diagnostic Observation Schedule (ADOS). ADOS's objective is to estimate the level of social deficiency in children which takes up to two hours (four 30 minutes sessions) and requires expert skills to conduct [Lord et al., 2000]. The Long time of the exam and its requirement for expert knowledge prevent its widespread application for early-stage screenings. In recent years researchers have developed machine learning and computer vision algorithms to automatically identify ASD behaviors based on video recordings in both controlled and uncontrolled settings [Chen and Zhao, 2019, Li et al., 2018, Li et al., 2019, Zunino et al., 2018, Li et al., 2020]. Some studies use subjects' gaze patterns [Wang et al., 2015b, Jiang and Zhao, 2017], interpretation of facial expressions and facial emotions [Liu et al., 2016, Tanaka and Sung, 2016], and gestures [Anzulewicz et al., 2016] to diagnose ASD. In [Wang et al., 2015b] the authors classify ASD patients based on their gaze patterns using a support

vector machines (SVM) classifier. As ASD patients have difficulty in recognizing faces and facial expressions, [Liu et al., 2016] investigates patterns of scanning facial markers by the subjects to identify those with potential ASD. [Anzulewicz et al., 2016] employed a touch-sensitive tablet with an embedded motion sensor to capture kinematic gestures performed by children while playing games. They were able to identify ASD with 93% accuracy among the participants. Most of these methods disregard the useful perceptual, temporal, and motion information from body parts used in action recognition techniques. Despite being effective in several scenarios, they can fail in many uncontrolled settings especially when the face is not visible or the quality of the captured video is not high enough to interpret the expressions.

The behavior of ASD children can be characterized by their defective interpersonal communication. A child with ASD may not respond to someone calling his/her name in the form of eye contact or head orientation [Liu et al., 2017]. [Rehg et al., 2013] addresses the challenges related to the communicative activities through parsing of subject's interactions into their atomic components. Therefore, it is able to rate the level of social dyadic interactions and evaluate the quality of a child's engagement. A public dataset called Multi-Modal Dyadic Behavior (MMDB) is recorded and released within the context of this study. [Marinoiu et al., 2018] proposes a setting for robot-assisted therapy of ASD children. In every session, the subject is accompanied by a robot while performing physical activities. Their multimodal dataset (DE-ENIGMA) consists of RGB and depth recordings which are utilized to evaluate actions as well as emotional expressions of children.

Action recognition methods can utilize appearance and motion information as well as articulated pose structures to model human behavior more accurately. Recently, various studies targeted the diagnosis of ASD by analyzing behavioral cues through action detection [Tian et al., 2019, Negin et al., 2021]. They are mainly focused on detecting repetitive behaviors (e.g. stereotype, self-injurious, ritualistic behaviors) that are considered as an important indicative of ASD [Edition et al., 2013]. O-GAD is a temporal convolutional network [Tian et al., 2019] that directly predicts ASD from arbitrarily long videos by detecting atypical repetitive behaviors. Potentially, action recognition and detection techniques can be utilized for ASD video content analysis. Two-stream networks [Simonyan and Zisserman, 2014, Wang et al., 2015a] apply a 2D convolutional operation on individual frames and optical flow field to learn spatial and temporal features. To learn spatiotemporal features directly from raw video data, the C3D network [Tran et al., 2015] employs 3D convolution. Recurrent Neural Networks (RNNs) have been widely used for temporal action detection in untrimmed videos [Buch et al., 2017, Ma et al., 2016]. However, RNNs have difficulty detecting long-range temporal dependencies. Temporal Convolutional Network (TCN) [Lea et al., 2016] proposed as an alternative to resolve this problem and perform temporal action segmentation. Instead, our framework uses inflated inception modules to extract spatial and temporal features from RGB and flow information and fuses them to perform action recognition.

2 Methodology

In this section we discuss the developed deep learning based methods for the evaluation of action recognition tasks on our collected dataset. These methods leverage different modalities to produce baseline results for further research in the future. Based on the modality, these methods are generally defined as RGB-based, or Flow-based methods. We are more focused on deep-neural-network-based methods to narrow down our approach. The overall architecture of the framework is given in Fig 1. As we will discuss in Section 3.1, in the studied dataset, usually two or more people are performing the actions. Therefore, it is crucial to detect and track the target children in the videos. The proposed methods comprise three crucial steps (person detection, tracking and action recognition). Prior to the action recognition task, we extract the region of interests (ROIs) from the video frames (in our case the child). This is done by a person detection framework followed by a tracking algorithm. All these steps are done at the preprocessing level which we explain next.

2.1 Pre-processing

Person detection: the videos are based on human-human or human-object interactions. Hence, the aim is not only to detect but to track the subjects ('child' and 'subject' are used interchangeably) as well. For the first part, three different person detection algorithms are fine-tuned and utilized. **HOG** [Zhou, 2014] descriptor, a gradient-based object detector is applied to detect people at each frame. the gradients are extracted by dividing the image into 8 by 8 cells, where the feature histograms are calculated accordingly. Prior to detection the histogram gradients are normalized. **YOLOv3** [Redmon and Farhadi, 2018], a strong end-to-end real-time object detection net-

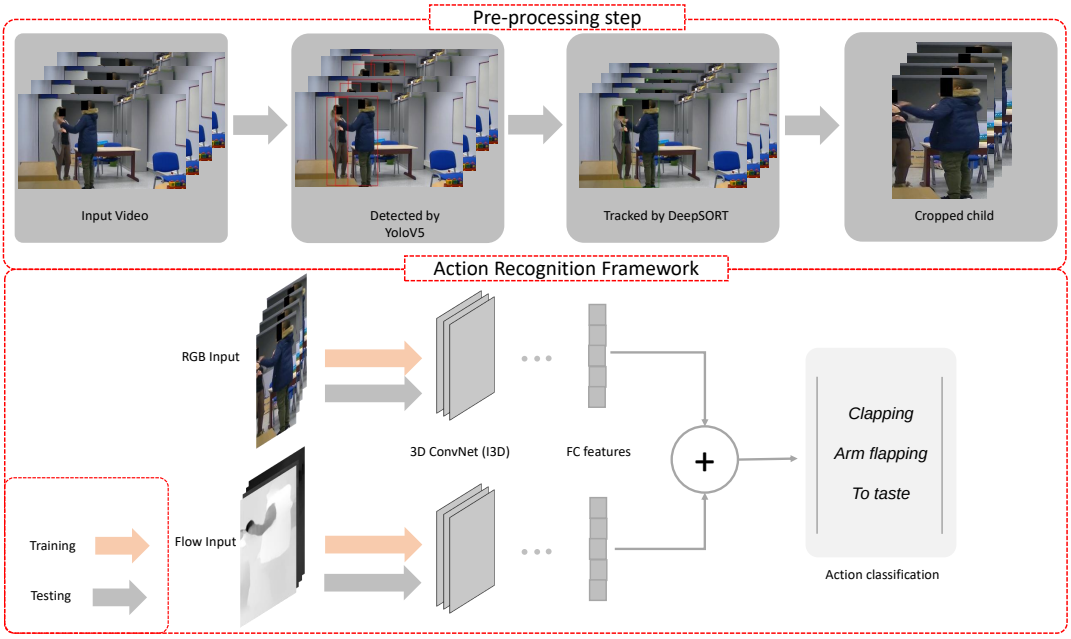


Figure 1: Action recognition pipeline based on I3D framework. The top box the necessary pre-processing required to extract the desired subject from the whole video. The box below is framework, in which a feature vector of 4096 in a Fully Connected (FC) layer achieved from a pre-trained I3D using RGB and flow inputs, individually. The features are concatenated at the last layer to predict desired action.

work, is used for person detection. This network consists of 3 x 3 and 1 x 1 convolutional and residual layers. It divides the image into a grid of S x S with each grid predicting anchor boxes B (for further details in [Redmon and Farhadi, 2018]). A model pre-trained on COCO [Liu et al., 2017] having 80 classes is used for our experiments.

YOLOv3 struggled at occlusion points in our dataset making it harder for the tracking algorithm to work properly. Therefore, experiments were carried out with **YOLOv5**, an improved successor of YOLOv3 with better accuracy for occluded objects, achieving desired results as illustrated in Fig 2.

Tracking: in order to track the desired target subject in the videos, the **SORT** which is an effective real-time tracking algorithm based on Kalman filter and Hungarian algorithm, is utilized. However, the SORT algorithm struggles with ID switching issue. Consequently, the tracking also performed with two other strong tracking algorithms [Zhang et al., 2020, Wojke et al., 2017]. **FairMOT** [Zhang et al., 2020], is an anchor free state-of-the-art tracking and detection algorithm based on an encoder-decoder network to extract high resolution features from an image. We utilize a pre-trained model based on MOT17 [Milan et al., 2016] and CrowdHuman [Shao et al., 2018] in our evaluations. The end results of the model is shown in the Fig 3. The obtained results were not

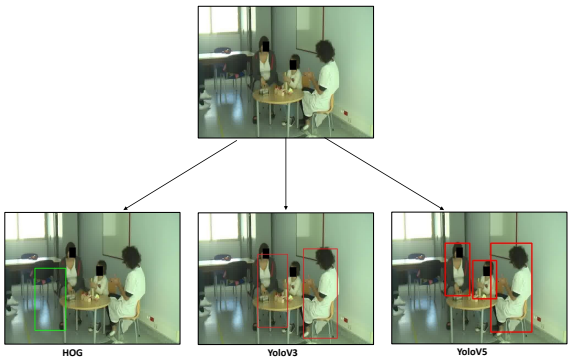


Figure 2: Examples of person detection algorithms applied: using HOG (Left), Yolov3 (Center), and Yolov5 (Right) detectors

good enough to be used for our action recognition task. The reason could be evaluating the dataset on the pre-trained model without training. We did not train the object detector and RE-ID module of this method specifically for our dataset due to lack of annotated data. To avoid training process the **DeepSORT** [Wojke et al., 2017] algorithm is employed. Deepsort works similarly to that of SORT algorithm including the Kalman filter for tracking the missing tracks and the Hungarian algorithm with a deep appearance descriptor to handle occlusion and view-point changes. The DeepSORT accuracy highly rely on object detec-

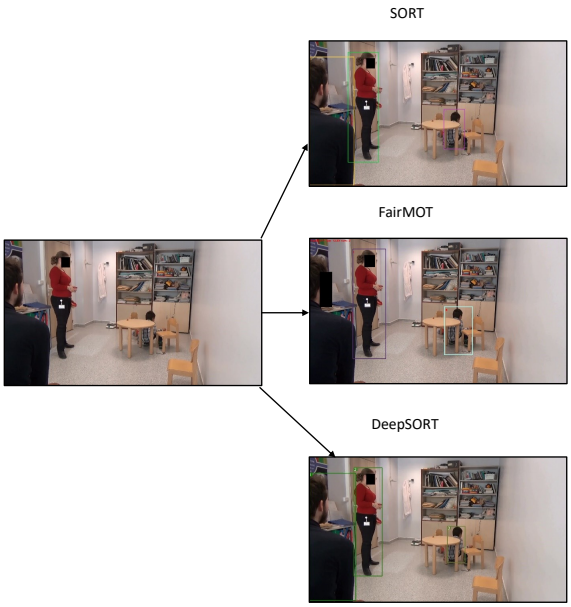


Figure 3: Tracking results using SORT (Top), FairMOT (middle), and DeepSORT (bottom) algorithms

tor module performance. DeepSORT while combined with YOLOv5 pre-trained on MOT17 and COCO datasets, separately achieved the desired results when applied on our datasets (Fig 3).

Extraction of Region of Interest (ROI): the recorded videos in our dataset are based on human-human interaction, where the ASD actions only appear in the child's actions. Therefore, after detection and tracking, the targeted subject is cropped at each frame using the assigned tracking ID and detected bounding boxes. As the subjects are from different age ranges, their body ratios varies. The cropped bounding boxes are resized to a fixed size by keeping height to width ratios unchanged to keep the evaluations independent from the body ratios.

Optical flow extraction: the optical flow is extracted by utilizing the **RAFT** (Recurrent All Pairs Field Transformers for Optical Flow) network [Teed and Deng, 2020]. The RAFT architecture extracts per pixel feature using an encoder, then it generates 4D correlation volumes for all pairs of pixels from all frames. A recurrent unit looks up the correlation volumes to update the flow field, iteratively. A pre-trained model is used to extract the optical flow. The normalized grayscale horizontal and vertical optical flow images are combined at channel dimension before feeding to the 3D convNets.

2.2 Two-stream Fusion Network

In this section, we discuss the deep-learning-based architecture for ASD actions in children. We study the effect of the RGB and flow modalities and later on, we perform a fusion of both at different levels of the network. Inception-VI is used as the backbone network by inflating all the 2D filters and pooling kernels to 3D. Thus, a 2D square filter of N x N becomes cubic N x N x N to include the temporal dimension.

The Inception-VI base architecture consists of a first convolutional layer with a stride 2, then four max-pooling layers having a stride of 2, and a 7 x 7 average-pooling layer leading to the last linear classification layer, in addition to the max-pooling layers in the parallel Inception branches. There was no temporal pooling applied in the first two max-pooling layers utilizing 1 x 3 x 3 kernels and stride 1 in time. While symmetrical kernels and strides were employed in all other max-pooling layers. The model is trained with 64-frames snippets but a whole video was used during testing by averaging the predictions temporally.

Fusion: in our dataset, the ASD actions include abrupt motions of the body parts. These motion cues are more difficult to capture by RGB but plays a significant role in analyzing such actions. Using motion information from optical flow can help to achieve better representations and recognition of these type of actions. Therefore, both RGB and optical flow streams were trained and tested separately prior to fusion to identify their contribution towards recognizing the action successfully. Later on, we perform a fusion of both modalities at different levels.

The early fusion variant does not fuse the raw data but instead the RGB and optical flow modalities are feed to the first convolution of the I3D network to extract initial features. These features from both modalities are then concatenated. The rest of the network remained unchanged obtaining a single stream architecture for both modalities.

Fusion at test time is a situation where both modalities are trained separately. The fusion occurs only during test time by averaging the results from both modalities.

A late fusion scenario was adopted by extracting features from the last layers of two-streams of I3D having RGB and optical flow inputs and concatenating them at the last layer. An input of the size $b \times 64 \times c \times w \times h$, with b being the batch size, c the input channel, w and h are the width and height of the each image, is provided to the model. The c should be 3 for RGB and 2 for optical flow modality. Feature vectors of size 4096 are extracted from the last Fully Connected (FC) layer of both RGB and optical flow

streams, separately. These features were fused at this point by a concatenation operation. An output equal to the number of classes resulted for classification.

3 Experiments

In this section, we first explain the collected dataset in addition to the augmentation techniques we utilized to increase size of the dataset. Next, we describe the experimental details used for model training and finally, we report the results and ablation studies.

Figure 4: Instance frames of the vidoes in the collected dataset. From top row to down: "clapping", "arm flapping", "to taste", "jump-up", and "others"

3.1 Activis dataset

The Activis dataset consists of 60 children recorded during child assessment sessions with presence of clinicians at the hospital. Their age ranges from 3 - 6 years. A trained clinician assessed the child based on communication skills, social interaction, and their imaginative use of materials according to the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) tool. Each session lasts for 50 minutes, the videos were recorded in a natural environment. Each child was diagnosed with a possible Autism disorder during different ADOS activities like *anniversary, playing with bubbles, construction, joint game* etc. The untrimmed videos were labeled by a psychologist for a possible 1) ADOS activity, and 2) repetitive action that was a potential indication of ASD disorder. The ADOS activity-based action is out of this paper's scope and will be discussed in future work. The actions in the dataset were divided to five categories (388 videos): *arm-flapping, clapping, to-taste, jump-up and others* (Fig 4). Each untrimmed video was clipped to 2 - 20 seconds clips based on the action. Blurred, distorted, and out-of-frame videos were discarded achieving a total of 388 trimmed videos. An HD camera was used for recording having 30 fps. With this conversion, there are a total of 12.5K frames in the dataset. Some of the subjects did not perform the same action, therefore, the data is not subject-oriented and highly imbalanced. More detailed information about the dataset is described in Table 1.

Table 1: Detailed information about the number of videos and frames of each action in the Activis dataset.

Action	**# of videos**	**Min/max/avg frames**	**Total no. of frames**
Clap	33	65/95/208	2447
Arm-flapping	59	58/112/198	6082
To-taste	132	73/157/388	18824
Jump-up	29	45/110/260	2270
Others	134	76/264/650	95695

3.2 SSBD dataset

Self-Stimulatory Behaviour Dataset (SSBD) is a publicly available dataset proposed in [Rajagopalan et al., 2013]. The data has been collected from different online portals, such as Youtube, Vimeo, and Dailymotion etc. The data consists of three distant autistic actions including *Armflapping, Headbanging, and Spinning*. The total actual data reported in original paper was 75 videos, in which only 66 are downloadable due to privacy concerns of Youtube. We validate our methods on this dataset along with our own dataset.

3.3 Data Augmentation

To increase the dataset and help the model converge effectively, three different augmentations are introduced: Horizontal flipping, up-sampling, and down-sampling. First, the videos were flipped horizontally covering the mirror effect and thus increasing the data. Next, temporal augmentations were also considered. Each video is up-sampled with a factor of 1.5 to

increase the number of frames within a video. Finally, we down-sample the videos with a factor of 0.5, thus throwing away some information from the video clip. These augmentations helped in generalizing the network. We compare the results before and after the augmentation and noted in the Table 2.

3.4 Experimental details

A two-stream I3D model pre-trained on the Kinetics dataset for both RGB and optical flow modalities was used. Initially, we train both streams separately without freezing any layers, at this point the model did not converge on either of the modality. A possible reason was that the dataset was small as compared to the network depth. Therefore, we used another approach, this time we froze the model and only trained the last layer. This fine-tuning technique helped the model in converging. We did the same for both streams of RGB and optical flow and noted the results separately. Secondly, we joined the two modalities during testing via a softmax function. But the results did not improve as expected. Lastly, for the fusion scenario, we extracted features from the pre-trained network, all layers being frozen, and concatenate them at the last layer. This concatenation helped in increasing the accuracy.

The model was trained for 100 epochs with a batch size of 16. An SGD optimizer was used with momentum set to 0.9 in all cases. A multi-GPU scenario was adapted by utilizing 4 GPUs for training the model. We started with a learning rate of 0.01 and fine-tuned its hyperparameters on the validation set of the proposed dataset. The complete architecture was implemented in Pytorch on a Linux operating system.

Table 2: Results of different methods on the collected dataset. Results are stated on augmented data for all networks. A late fusion of two-stream I3D (RGB and optical flow) resulted in higher accuracy (in bold). All the methods use a 5-fold-cross-validation technique.

Method	**Pre-trained (Kinetics)**	**Acc. no Aug**	**Acc. with Aug.**	**F1 Score Macro/weighted**
HOG + SVM	No	23.5	22.7	15/18
CSN rgb	Yes (IG-65m)	80.31	81	63/71
X3D rgb	Yes	85.3	85.76	68/77
I3D Flow	Yes	78.81	80.30	61/70
I3D rgb	Yes	84.97	85.01	71/79
Fusion (rgb+flow)	**Yes**	**85.6**	**86.04**	**72/81**

3.5 Results and Discussion

Cross-validation is a re-sampling technique used to evaluate architectures on limited data. Since our dataset is small and highly imbalance, we adopt StratifiedK-fold approach of Sklearn library. StratifiedK-folds is the most usable cross-validation method, where the data is divided into k folds keeping a balanced ratio of each class. In our experiments, we keep the value of k 5, generating 5-folds to train the model and average their results. The data is divided into 5-folds and tested for low bias and variance. We proposed the baseline results on Activis and the SSBD datasets using different modalities and their fusion. We also perform experiments with other methods including X3D [Feichtenhofer, 2020], CSN [Tran et al., 2019] and a non deep neural network method: HOG + SVM [Cristianini and Ricci, 2008] classifier. A comparison is shown Table 2, and 3.

Furthermore, we provide a confusion matrix of different methods on Activis dataset in Figure 5 to better understand the efficiency of our approach. The confusion matrix results are averaged of the 5-folds. We also detailed the confusion matrix results on SSBD dataset in Figure 6.

Figure 5: Confusion matrix of proposed method on Activis dataset.

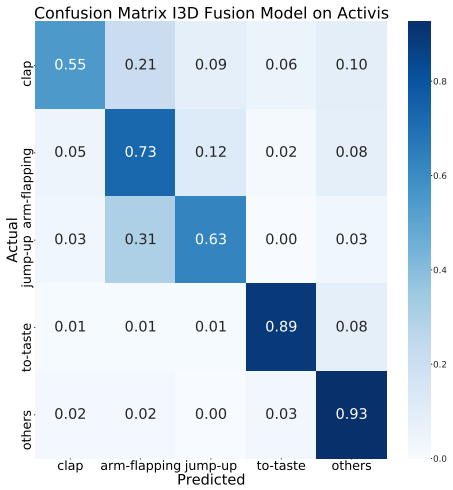
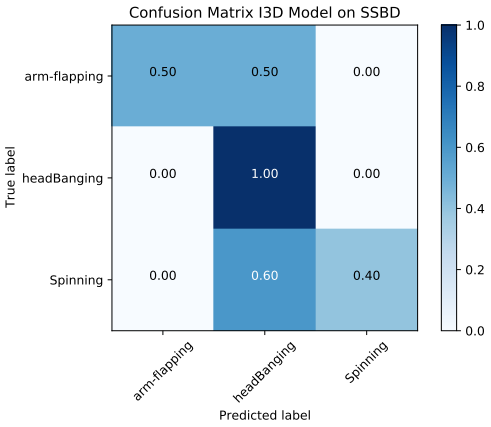


Table 3: Results of different methods on SSBD dataset.

Method	**Pre-trained (Kinetics)**	**Acc.**	**F1 Score macro/weighted**
CSN RGB	Yes (IG-65m)	71.11	40/59
X3D RGB	Yes	64.98	45/61
I3D Flow	Yes	30.97	36/48
I3D RGB	Yes	76.92	55/60
Fusion (RGB + Flow)	**Yes**	**75.62**	**62.5/69**

Figure 6: Confusion matrix of proposed method on SSBD dataset.



3.6 Ablation Study

In table 4 we summarize the ablation study in our experiments. We study the effect of each stream by training the RGB and optical flow branches separately. We also analyze the fusion of the two modalities at different levels (early and late fusion). It is interesting to see that the fusion at last layer during training is more effective compared to the fusion at the test time.

Table 4: Ablation study of I3D and its variants.

Method	**Acc. (%)**
RGB only	78.8
Flow only	84.97
Fusing during test	82.1
Early fusion	83.3
Late fusion	**85.6**

4 Conclusion

In this paper, we attempt to apply computer vision techniques for the diagnosis of ASD behaviors. For this purpose, we collected and annotated a rich dataset of children's stereotypic behavior videos recorded in an uncontrolled environment during their actual diagnosis of ASD. We developed several action-recognition-based frameworks to recognize these characteristic behaviors. Several experiments with different modalities and networks are performed to propose a baseline for action recognition of children having ASD. The baseline results show that a late fusion of the I3D network having two modalities outperforms the other methods.

In the future, we plan to increase our dataset by annotating more videos with more action classes. In addition, we want to build a subject-oriented dataset for long-term action tasks, for the development of technology that will be useful to parents/ caregivers and clinicians for early diagnosis.

REFERENCES

(2021). World Health Organization (WHO): Autism spectrum disorders Key Facts. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.

Anzulewicz, A., Sobota, K., and Delafield-Butt, J. T. (2016). Toward the autism motor signature: Gesture patterns during smart tablet gameplay identify children with autism. *Scientific reports*, 6(1):1–13.

Buch, S., Escorcia, V., Shen, C., Ghanem, B., and Carlos Niebles, J. (2017). Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920.

Chen, S. and Zhao, Q. (2019). Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1181–1190.

Cristianini, N. and Ricci, E. (2008). *Support Vector Machines*, pages 928–932. Springer US, Boston, MA.

de Belen, R. A. J., Bednarz, T., Sowmya, A., and Del Favero, D. (2020). Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):1–20.

Edition, F. et al. (2013). Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21.

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213.

Hashemi, J., Spina, T. V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., and Sapiro, G. (2012). A computer vision approach for the assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE.

Huerta, M. and Lord, C. (2012). Diagnostic evaluation of autism spectrum disorders. *Pediatric Clinics of North America*, 59(1):103.

Jiang, M. and Zhao, Q. (2017). Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the ieee international conference on computer vision*, pages 3267–3276.

Kanner, L. et al. (1943). Autistic disturbances of affective contact. *Nervous child*, 2(3):217–250.

Knopf, A. (2020). Autism prevalence increases from 1 in 60 to 1 in 54: Cdc. *The Brown University Child and Adolescent Behavior Letter*, 36(6):4–4.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer.

Lewis, M. H. and Bodfish, J. W. (1998). Repetitive behavior disorders in autism. *Mental retardation and developmental disabilities research reviews*, 4(2):80–89.

Li, B., Mehta, S., Aneja, D., Foster, C., Ventola, P., Shic, F., and Shapiro, L. (2019). A facial affect analysis system for autism spectrum disorder. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4549–4553. IEEE.

Li, J., Zhong, Y., Han, J., Ouyang, G., Li, X., and Liu, H. (2020). Classifying asd children with lstm based on raw videos. *Neurocomputing*, 390:226–238.

Li, J., Zhong, Y., and Ouyang, G. (2018). Identification of asd children based on video data. In *2018 24th International conference on pattern recognition (ICPR)*, pages 367–372. IEEE.

Liu, W., Li, M., and Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8):888–898.

Liu, W., Zhou, T., Zhang, C., Zou, X., and Li, M. (2017). Response to name: A dataset and a multimodal machine learning framework towards autism study. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 178–183. IEEE.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3):205–223.

Ma, S., Sigal, L., and Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1942–1950.

Marinoiu, E., Zanfir, M., Olaru, V., and Sminchisescu, C. (2018). 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2158–2167.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., and Ozyer, G. T. (2021). Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. *Neurocomputing*, 446:145–155.

O'Roak, B. J. and State, M. W. (2008). Autism genetics: strategies, challenges, and opportunities. *Autism Research*, 1(1):4–17.

Pandey, P., Prathosh, A., Kohli, M., and Pritchard, J. (2020). Guided weak supervision for action recognition with scarce data to assess skills of children with autism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 463–470.

Rajagopalan, S., Dhall, A., and Goecke, R. (2013). Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 755–761.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al. (2013). Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. (2018). Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*.

Tanaka, J. W. and Sung, A. (2016). The "eye avoidance" hypothesis of autism face processing. *Journal of autism and developmental disorders*, 46(5):1538–1552.

Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer.

Tian, Y., Min, X., Zhai, G., and Gao, Z. (2019). Video-based early asd detection via temporal pyramid networks. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 272–277. IEEE.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Tran, D., Wang, H., Torresani, L., and Feiszli, M. (2019). Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561.

Wang, L., Xiong, Y., Wang, Z., and Qiao, Y. (2015a). Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*.

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., and Zhao, Q. (2015b). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE.

Zhang, Y., Tian, Y., Wu, P., and Chen, D. (2021). Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder. *Sensors*, 21(2):411.

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2020). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*.

Zhou, D. (2014). *Real-time animal detection system for intelligent vehicles*. PhD thesis, Université d'Ottawa/University of Ottawa.

Zunino, A., Morerio, P., Cavallo, A., Ansuini, C., Podda, J., Battaglia, F., Veneselli, E., Becchio, C., and Murino, V. (2018). Video gesture analysis for autism spectrum disorder detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3421–3426. IEEE.