Automatic Detection of ADHD and ASD from Expressive Behaviour in RGBD Data

Shashank Jaiswal[1] Michel F. Valstar[1] Alinda Gillott[2] David Daley[3]
[1]School of Computer Science, The University of Nottingham
[2]Nottingham City Asperger Service & ADHD Clinic
[3]Institute of Mental Health, The University of Nottingham

***Abstract*— Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD) are neurodevelopmental conditions which impact on a significant number of children and adults. Currently, the diagnosis of such disorders is done by experts who employ standard questionnaires and look for certain behavioural markers through manual observation. Such methods for their diagnosis are not only subjective, difficult to repeat, and costly but also extremely time consuming. In this work, we present a novel methodology to aid diagnostic predictions about the presence/absence of ADHD and ASD by automatic visual analysis of a persons behaviour. To do so, we conduct the questionnaires in a computer-mediated way while recording participants with modern RGBD (Colour+Depth) sensors. In contrast to previous automatic approaches which have focussed only on detecting certain behavioural markers, our approach provides a fully automatic end-to-end system to directly predict ADHD and ASD in adults. Using state of the art facial expression analysis based on Dynamic Deep Learning and 3D analysis of behaviour, we attain classification rates of 96% for Controls vs Condition (ADHD/ASD) groups and 94% for Comorbid (ADHD+ASD) vs ASD only group. We show that our system is a potentially useful time saving contribution to the clinical diagnosis of ADHD and ASD.**

I. INTRODUCTION

The last 5 years have seen a steady progress in automatic expressive behaviour analysis, with the detection and tracking of faces [1][2][3][4], recognition of facial muscle actions [5][6][7], and accurate head pose estimation [1][8] all now possible under mild environmental constraints. This has renewed the interest of researchers to employ such behaviour analysis in the medical domain, targeting so-called *behaviomedical* conditions that alter one's expressive behaviour [9]. In this paper we use state of the art facial expression analysis and RGBD head motion analysis to help in the diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD).

ADHD is a neurodevelopmental condition affecting a large number of people and it has been estimated that at least 2.5% of the general adult population is affected by it [10]. ADHD is characterized by symptoms such as hyperactivity, impulsivity, inattention, etc. [11][12]. It usually begins in early childhood and quite often the symptoms persists into adulthood [13]. It is widely believed that both genetic [14] and environmental influences [15] contribute to the underlying cause of this disorder. Presently, the diagnosis of ADHD is made following the criteria of the DSM-5 [16],

978-1-5090-4023-0/17/$31.00

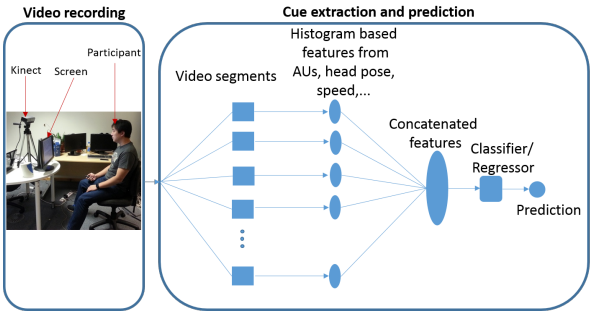

Fig. 1: Overview of our system. A participant follows instructions on a screen while being recorded by a Kinect 2 camera. Deep Learning and RGB-D behaviour analysis of each video segment leads to successful ASD/ADHD classification.

which involve mechanisms to validate hyperactivity, attention deficit and impulsivity. The diagnosis is made by experts using a combination of developmental history, collateral information, psychometrics and behavioural observation and impairment. This is often difficult and time consuming.

ADHD is also known to show co-morbidity with ASD (Autism Spectrum Disorder). ASD is a neuro-developmental condition which is characterised by impairments in social interaction and communication and restricted, repetitive or stereotyped behaviours and interests. It has been found that a significant number of people with ASD also show symptoms of ADHD [17]. Treatment methods also vary for all 3 groups of people i.e. only ASD, only ADHD, and comorbid (ADHD+ASD). Hence, accurate diagnosis can have important implications for treatment. However, currently the manual diagnosis for each of these disorders has to be done separately which requires more time.

Although there has been a lot of research in the area of ADHD and ASD and their diagnosis using brain scanners and manual observation of subjects for extended periods of time by psychological experts, there has been relatively little work in the direction of developing automated diagnostic aids for ADHD and ASD using easily available devices (e.g. video camera). The current methods of diagnosis are not only time consuming but they are also susceptible to human decision making bias. Development of machine learning methods which can be used as a tool for decision making by

human experts, could not only save time but will also help in bringing more objective, repeatable measures in the decision making process.

Currently available commercial systems (e.g. QbTest [18]) seek to automate the process of ADHD diagnosis uses only head motion of a person as a proxy for the activity of the subject. The other aspects of head actions including its pose are not taken into account directly. The head motion itself is captured using a normal 2D imaging camera which has limited ability to capture motion in 3D. Facial expression is another aspect which is completely ignored in current ADHD assessment systems. Facial expressions and gestures can provide important cues about the psychological state of a person. There has been some work which indicates that facial expressions could be useful in the diagnosis of certain psychological disorders[19], [20]. But to the best of our knowledge, until now there has been no research which establishes the relationship between facial expressions and ADHD/ASD.

In this work we aim to make the diagnostic procedure for ADHD and ASD easier, more efficient and more objective through automatic analysis of a person's behaviour. We propose a computer-vision based approach to automatically aid diagnosis of ADHD and ASD. We extract high level features from tracked faces in videos to learn classification models for ADHD and ASD prediction. We adapt a recently proposed Dynamic Deep Learning method to recognise facial action units from RGB data [21], and use face tracking data from RGBD (colour+depth) images recorded using a Kinect 2.0 sensor camera to obtain head actions and facial animation unit parameters.

We also present a first of its kind RGBD database in which 55 subjects who have previously been diagnosed with ADHD or ASD as well as subjects from a healthy control group were recorded in a controlled setting. We evaluate our proposed approach on this database and show that that our approach performs highly accurately on ADHD and ASD classification tasks achieving classification rates of 96% for Controls vs Condition (ADHD/ASD) group and 94% for Comorbid (ADHD+ASD) vs ASD only group.

In summary, our main contributions are:

- A novel fully automatic approach for making diagnostic predictions for ADHD and ASD directly from videos.
- Establishing the relationship between facial expression/gestures and neurodevelopmental conditions such as like ADHD and ASD.
- A new database for evaluating computer vision based algorithms on the task of predicting ADHD and ASD diagnosis.

II. Related Work

The field of using Computer vision techniques for monitoring people for ADHD and ASD is still in its infancy and there has been limited research reporting on this topic. Below we describe some of the existing works which aim towards automatic detection of certain markers which could help in the diagnosis of ADHD and ASD.

A. Detection for ADHD

Some preliminary studies have been conducted to demonstrate the use of depth capturing cameras to monitor the activities of people. For e.g. Hernandez-Vela et al. [22] extracted 3D skeletal model of human body, using RGB-D image sequences. Using this skeletal model, they tracked 14 reference points corresponding to skeletal joints and used them to detect certain body gestures often found in children having ADHD. For detecting such gestures, they used Dynamic Time Warping [23]. By measuring the similarity between a temporal sequence of images with a reference sequence of a gesture, they demonstrated that they can recognize a set of defined gestures related to ADHD indicators.

In [24], a system was developed for tracking people across multiple cameras and sensors. They used depth measuring cameras (Microsoft Kinect) to monitor the movement of children in a classroom setting. The authors used agglomerative hierarchical clustering to segment different objects and tracked different individuals using covariance descriptors. One of the applications they proposed for such a system would be to record the motion tracks and velocity profiles of people, to measure their activity level.

QbTest [18] is one of the most successful commercially available systems for monitoring and diagnosis of ADHD. It measures 3 main indicators of ADHD: hyperactivity, inattention and impulsivity. It combines head motion tracking with a computer based test. The head motion tracking is designed to measure the hyperactivity of the subject. For this purpose, the subject taking the test is required to wear a head band which has a reflector attached to it. The camera in front of the subject, tracks the movement of the reflector. However, the system's ability to capture the full facial information is limited as it does not track the entire face thus ignoring the 3D head pose and facial expression information. To measure the inattention and impulsivity, the subject has to take a computerised continuous performance test in which the participant has to respond quickly and accurately to certain geometrical shapes displayed on the screen. The whole test lasts for 15-20 minutes and the head motion is tracked during the entire time. After the test, the result is compared to the norm data corresponding to the subject's age and gender and a report is generated for assessment by clinicians.

B. Detection for ASD

One of the pioneering works in the field of ASD diagnosis was done by Hashemi et al. [25]. In this work, the authors developed computer vision based methods to identify certain behavioural markers based on Autism Observation Scale for Infants (AOSI) related to visual attention and motor patterns. For assessing visual attention, they focused on 3 main behavioural markers, namely sharing interest, visual tracking and disengagement of attention. These behavioural markers were detected by estimating the head pose in the left-right direction (yaw) and in the up-down direction (pitch). The head pose was estimated by tracking the position of certain facial features (eyes, nose, ear, etc.).

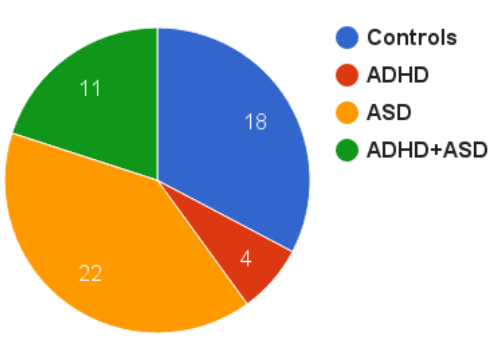

Fig. 2: Distribution of participants in KOMAA dataset.

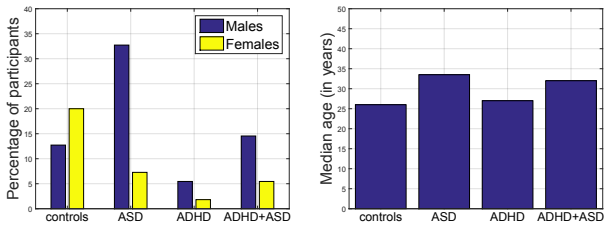

Fig. 3: Gender distribution and median age of participants within different groups in the KOMAA dataset.

In [26], the authors presented another computer vision based approach for studying autism by retrieving social games and other forms of social interactions between adults and children in videos. They proposed to do this by defining social games as quasi-periodic spatio-temporal patterns. In order to retrieve such patterns from unstructured videos, the authors represent each frame using a histogram of spatio-temporal words derived from space-time interest points. The frames are clustered based on their histograms to represent the video as a sequence of cluster (keyframes) labels. The quasi-periodic pattern is found by searching for co-occurrences of these keyframe labels in time.

In [27], the authors proposed an algorithm for detecting self-stimulatory behaviour which is a common behavioural marker in individuals with autism. They computed motion descriptor using dominant motion flow in the tracked body regions, to build a model for detecting self-stimulatory behaviour in videos. Similarly, in [28], the authors measure children's engagement level in social interactions using low level optical flow based features.

Most of the above mentioned works have concentrated on detecting certain pre-defined behavioural markers which are often associated with either ADHD or ASD in children. They are preliminary works whose effectiveness in predicting the actual ADHD and ASD diagnosis still remains to be seen. On the other hand, this work poses the diagnosis of ADHD and ASD, directly as a machine learning problem. Our work is one of the first which attempts to learn models for directly predicting conditions such as ADHD and ASD using high level facial features which can be reliably computed nowadays. This work also differs from other works in the sense that it is mainly focusses on ADHD and ASD diagnosis in adults rather than children.

III. Data Collection

We collected a dataset 'KOMAA' (Kinect Data for Objective Measurement of ADHD and ASD) for the purpose of evaluating our proposed method. The database consists of video recordings from a total of 55 subjects. The length of each video is approximately 12 min. and is recorded using a Kinect 2.0 device which is capable of capturing high resolution RGB and depth images. All the participants in the recording were adults over the age of 18 years. During the recordings the subjects sit in front of a computer screen and have to read and listen to a set of 12 short stories. Each story is accompanied by 2-3 questions which the subjects have to answer in their voice. These stories have been selected from the 'Strange Stories' task [29] which is often used as psychological test for the diagnosis of ASD. The text of each story along with the corresponding questions were displayed on the screen. Additionally, a pre-recorded voice was played reading out the story and the corresponding questions. Such a setup was prepared so as to simulate the effect of an actual person telling the story and asking the questions, but at the same time keeping the setup as automated as possible.

The subjects in this database can be divided into four different categories. The first category is the control group which consists of subjects who show no symptoms of ADHD/ASD and have never been diagnosed with either ADHD or ASD. In order to make sure that the subjects recruited in the control group do not have any chance of having ADHD or ASD, each subject was asked to complete 2 screening questionnaires: Adult ADHD Self-Report Scale (ASRS) [30]) and Autism Spectrum Quotient (AQ10) [31]. The ASRS is a screening measure for ADHD symptoms in adults consisting of 18 items. It is based on DSM-IV items for ADHD and is considered to have excellent reliability and validity [32]. Similarly, AQ10 is a screening measure for Autism symptoms consisting of 10 items and is widely used to measure the degree to which an adult has autistic traits. Only those participants who scored less than a certain threshold value in each of the questionnaires, were selected as a part of the control group. The threshold values for ASRS (Part A) and AQ10 were set to be 4 and 6 respectively. The other three categories include the ASD group (consisting of subjects who have been diagnosed with ASD), ADHD group (subjects who have been diagnosed with ADHD) and ASD+ADHD group (subjects who have been diagnosed with both ADHD and ASD.

The total number of subjects recruited into each category is shown in Fig. 2. The gender distribution and the median age of the participants in each category is also shown in Fig. 3.

IV. Methodology

Training statistical machine learning based classifiers which can automatically differentiate between subjects with

ADHD/ASD from healthy controls, is a difficult problem. The problem becomes even more challenging when the number of training examples are small. Deep learning based approaches which directly use low level pixel information to learn high level semantics, currently provide state-of-the-art performance on a number of computer vision tasks. However, using low level information on the limited number of training examples in our case, can lead to severe overfitting.

Our approach to training the classifiers involves computing high level feature descriptors corresponding to facial expressions (facial AUs), head pose and motion, etc. To compute the feature descriptors, each video is first divided into 12 segments corresponding to the 12 stories that the participants have to read while they were recorded. This has been done manually, but could easily be automated given that the timing of the delivery of the stories is controlled by the researcher.

For each video segment, histogram based feature descriptors are computed separately using pre-trained classifiers/regressors that detect individual behavioural cues. Grouping these cues per story helps to preserve temporal information which would otherwise be lost if histograms would have been computed over all the frames in a video, at a small price of multiplying the dimensionality of our overall feature vector by a factor 12. The combined set of feature descriptors from all segments in a recording are used for used for training the ADHD/ASD classification models (See Fig. 1). Below we describe the main components of our approach in more detail.

A. Feature descriptors

Six different sets of features are computed from the recorded video of each subject. Most of the features are computed on a per-frame basis, which are then converted into multiple histograms where each histogram is computed over all the frames in a video segment. The feature descriptors used in our approach are described below:

1) Dynamic Deep Learned Facial Action Units:
Facial action units (AU) are movement of individual or group of facial muscles defined according to the Facial Action Coding System (FACS)[33]. Anatomically based descriptors of facial expressions, they can be a good representative of the emotional and mental state of a person and can encode a large number of social signals. Intensities for a set of 6 AUs (AU1, AU2, AU4, AU12, AU15, AU20) and occurrence for AU45 (blinks) were estimated for each frame in video. For this purpose we used AU models trained using a slightly modified version of the deep CNNs described in [21]. The network architecture used for this purpose is shown in Fig. 4. This network does not use Bi-directional Long Short Term Memory (BLSTM) used by the original work in [21]. Histograms of AU intensities was computed over all the frames in a video segment. One histogram was computed for each AU consisting of 10 bins each. For AU45, the frequency of its occurrence and the average duration of its activation were estimated in each video segment. The histograms of all AU intensities and the AU45 statistics were concatenated together resulting in a 62 dimensional AU vector F_{au}, for each video segment.

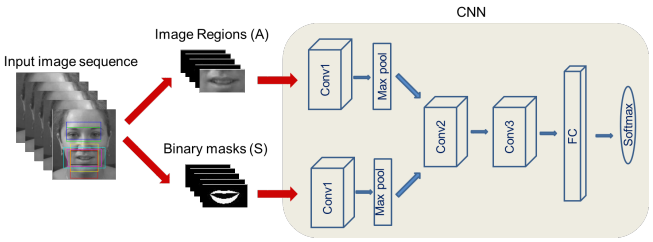

Fig. 4: Graphical overview of the CNN based approach used for predicting facial AUs [21].

2) Kinect Animation Units:
The Kinect also provides Animation Units (AnUs), geometry-based descriptors similar to mpeg-4 face animation parameters (FAPs)[34]. While they are not based on muscle actions and can not detect facial actions that only cause appearance changes, the fact that they are obtained from RGBD data makes them very reliable. The intensity of a number of AnUs were estimated for each frame in the video using the Kinect v2 library. In order to aggregate the statistics over each video segment, a histogram of ANU intensities was computed for each facial AnU. Each histogram consisted of 10 bins resulting in a 10 dimensional feature vector corresponding to each ANU. A total of 12 AnUs (6 corresponding to left and 6 to right part of the face) were used. The histograms from all 12 AnUs were concatenated, resulting in 120 dimensional AnU vector F_{an}, for each video segment.

3) Head Pose:
One of the major challenges for people with ADHD is their inability to do tasks which requires sustained attention. The pose of the head (in 3D space) can provide valuable cues about the attention state of a person at a certain instance of time. Since the participants in our study were required to complete the task by looking the computer screen, any deviation of the head pose away from the computer screen would indicate loss of attention.

The rotation of the head about the X, Y and Z axis (pitch, yaw and roll) were estimated for each frame of the video using the Kinect v2 software. The X, Y and Z axis are defined in reference to the location of the Kinect device as shown in Fig. 5. We assumed the median pose of the head to be the most attentive state. Rotation of the head away from the median pose were computed about the X, Y and Z axis separately. Histograms of these rotation angles were computed over the video segments for each axis separately. Each histogram consisted of 18 bins with equally spaced bin centres ranging from $-45°$ to $45°$. This resulted in a 54 dimensional head pose vector F_{hp}, for each video segment.

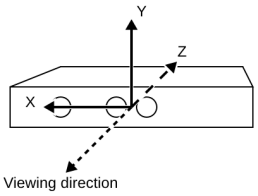

Fig. 5: Kinect coordinate system.

4) Speed of head movement:
Dynamics of head motion has been a less researched aspect in the field of psychological disorders. In order to investigate the role head motion, we estimated the speed of head motion at each frame of the video. For this purpose, we selected a set of stable facial landmarks belonging to eye corners and 4 points on the nose. The location of these stable facial landmarks are invariant to changes in facial expressions and hence suitable for estimating the motion of the head. The motion of the head is estimated by computing the location of the centroid C_i of the stable landmarks. The speed of head motion S_i at any frame i can be estimated by computing the displacement of the centroid as given below:

$$S_i = ||C_i - C_{i-1}|| * f \tag{1}$$

where f is the frame rate of the recorded video. In order to make speed estimation more reliable and invariant to any fluctuations in the frame rate, the estimated speed was smoothed by computing a moving average over 20 consecutive frames.

A histogram of the estimated speeds was computed to aggregate the statistics over each video segment. The histograms consists of 10 bins resulting in a 10 dimensional speed vector F_{sp}, for each video segment.

5) Cumulative Distance:
Hyperactivity is another major challenge associated with ADHD, implying that individuals with ADHD tend to display much higher levels of motoric behaviour than healthy individuals. The movement can be in the form of whole body movement or smaller movements confined to head (rotation) or hands and legs (fidgeting). To encode such information, the cumulative distance F_{cd} moved by the head during an entire video segment, was estimated by summing up the displacements of the centroid C_i given below:

$$F_{cd} = \sum_{i=1}^{n} ||C_i - C_{i-1}|| \tag{2}$$

where n is the total number of frames in the video segment.

6) Response Times:
The time taken to respond to each set of questions in the study was also used as features. Since there were 12 stories, each comprising a set of questions, a 12 dimensional response time vector F_{rt} was defined consisting of the response times (in seconds) for each set of questions.

B. Feature pre-processing and training models

Normalization: Each set of features (except the F_{rt}) were divided by the total number of frames in the video segment, to make them invariant to the length of video recording. The final set of features F was obtained by concatenating all sets of features (F_{au}, F_{an}, F_{hp}, F_{sp}, F_{cd}, F_{rt}) from all video segments. Each dimension in the resulting feature vector F is further normalized by computing the Z-scores.

Feature selection and training models Due to the high dimensionality of the resulting feature F compared to the number of training examples, any classifier trained directly on the entire feature-set is most likely to overfit the training data. In order to avoid such problem, a greedy forward feature selection was employed to capture the most relevant features and reduce the dimensionality. The classification models were trained using Support Vector Machines (SVM) with a Radial Basis function kernel.

V. Experiments and Discussion

Our approach was evaluated on the KOMAA dataset that we collected for this purpose from a total of 55 participants (see section III). The distribution of participants with ADHD, ASD and healthy controls is shown in Fig. 2.

To evaluate the performance of our approach in classifying each subject to the ASD, ADHD or the Control group, we followed a a 2 step procedure: In the first step we trained a classifier to distinguish between controls and condition group (participants diagnosed with either ADHD, ASD or both). In the second step, we trained another classifier to distinguish between ASD only group and Comorbid (ASD+ADHD) group. Since the ADHD only group was too small (only 4 participants), we did not had enough data to learn a robust classifier for this group.

TABLE I: Classification results for Controls vs Condition (ASD/ADHD) group.

Classifier	Correct	Incorrect
Controls	16	2
Condition	37	0

TABLE II: Classification results for Comorbid (ADHD+ASD) vs ASD group.

Classifier	Correct	Incorrect
Comorbid	9	2
ASD only	22	0

Our approach was evaluated using a leave-one-subject-out protocol, in which one subject is used for testing and the rest of the subjects are used for training. This process is repeated for each subject and the overall score is obtained by averaging over each test subject. The classification performance of our approach is shown in Table I and II. For classification into Control and Condition group, we obtain a very high classification accuracy of 96.4%. Similarly, for

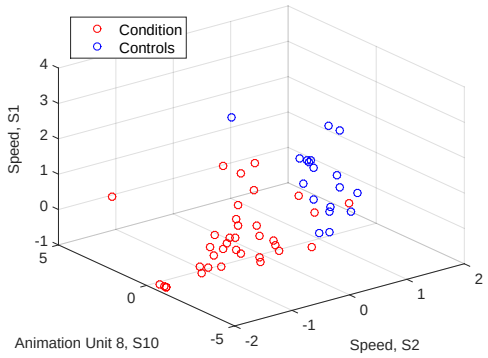

Fig. 6: Top 3 features distinguishing Condition (ASD/ADHD) from control group. Animation Unit 8 corresponds to lip-corner depressor. S1, S2, S10 denote video segments corresponding to story 1, 2 and 10 of the 'Strange Stories' task respectively.

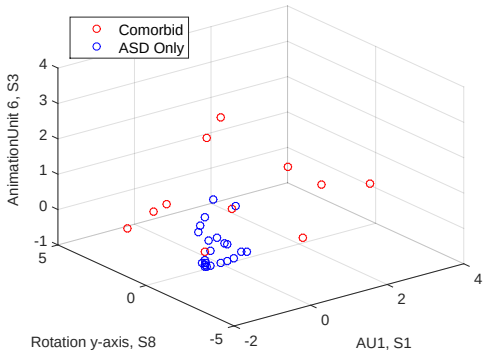

Fig. 7: Top 3 features distinguishing Comorbid (ASD+ADHD) from ASD only group. Animation Unit 6 and AU1 corresponds to lip-corner puller and inner-brow raiser respectively. S1, S3 and S8 denote video segments corresponding to story 1, 3 and 8 of the 'Strange Stories' task respectively.

classification into Comorbid(ASD+ADHD) and ASD only group, we obtain a high classification accuracy of 93.9%.

Looking at the individual contribution of different cues, Fig. 6 and 7, show the class separation provided by some of the important features selected by using the forward feature selection approach. From these figures, we can observe that for classification of Controls and Condition group, features such as Speed of head motion (from video segment corresponding to story 1 and 2) and Animation Unit 8 (lip-corner depressor from video segment corresponding to story 10 of 'Stange Stories task') were found to be most discriminative. For Comorbid vs ASD classification, AU1 (inner-brow raiser), AnU6 (lip-corner puller) and head rotation about Y-axis turn out be highly discriminative These features were extracted from the video segment corresponding to story 1, 3 and 8 of the Strange stories task respectively. In Fig. 8 and Fig. 9, we also show a list of top 30 features (for both classification problems) ranked according to their individual classification power.

VI. Conclusions

We presented a novel method for making diagnostic prediction of ADHD and ASD in test subjects through automatic video analysis. Facial cues such as head motion, facial expression and pose are used in learning models which can accurately predict ADHD and ASD. The role of facial expressions as a potential feature for classification of individuals with these disorders from healthy controls, was investigated. A high performance was achieved in terms of classification accuracy, which indicates a high potential for facial expressions and other facial gestures to be used for making automatic predictions for ADHD, ASD and other neurodevelopmental disorders.

Acknowledgements

The research reported in this paper was supported by the NIHR MindTech Healthcare Technology Co-operative. The work of Valstar is also supported by European Union Horizon 2020 research and innovation programme under grant agreement No 645378. The views represented are the views of the authors alone and do not necessarily represent the views of the Department of Health in England, NHS, or the National Institute for Health Research.

References

[1] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

[2] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

[3] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. of CVPR*, pages 532–539, Portland (OR), USA, 2013. IEEE.

[4] Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. In *European Conference on Computer Vision*, pages 645–661. Springer, 2016.

[5] Michel F Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2012.

[6] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[7] Michel Valstar, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeff Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *Proc. of FG*, Ljublijana, Slovenia, May 2015. IEEE.

[8] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083, 2016.

[9] Michel Valstar. Automatic behaviour understanding in medicine. In *Proc. of RFMIR, ICMI*, pages 57–60, Istambul, Turkey, November 2014. ACM.

[10] Viktória Simon, Pál Czobor, Sára Bálint, Ágnes Mészáros, and István Bitter. Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. *The British Journal of Psychiatry*, 194(3):204–211, 2009.

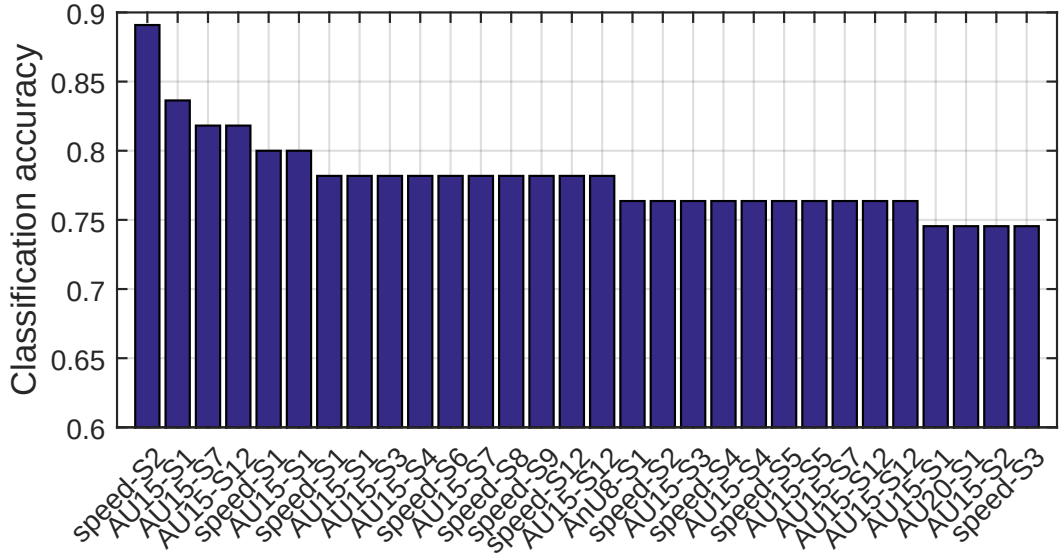

Fig. 8: Top 30 features for classification of Controls vs Condition group. Each feature is represented by its feature type followed by the video segment number it was computed on. For e.g. AU15-S1 means that the feature corresponds to AU15 intensity histogram computed from the video segment corresponding to story 1 of the 'Strange stories' task. Please note that the same feature name can appear more than once because they are different features coming from the same histogram.

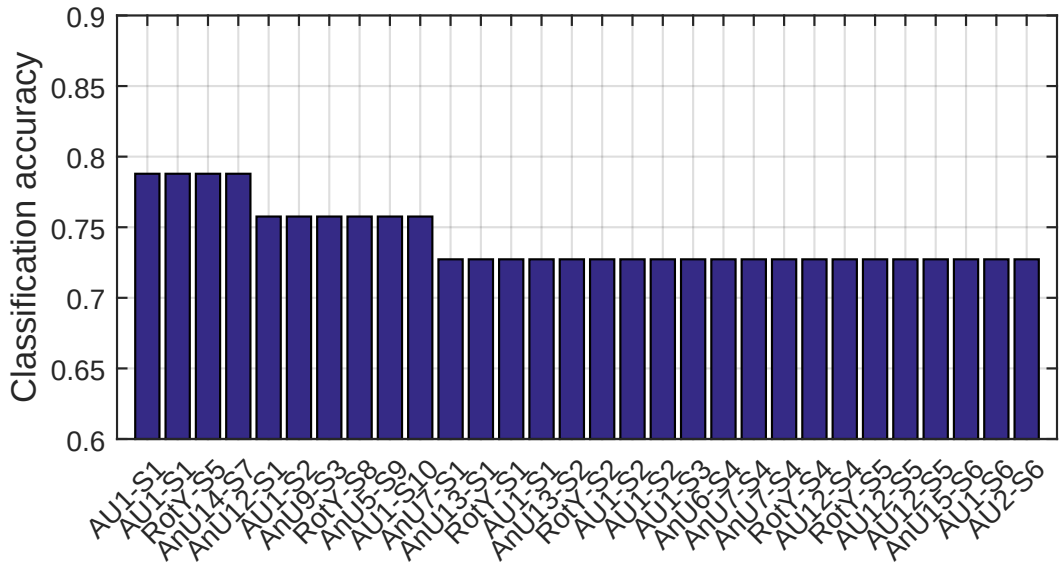

Fig. 9: Top 30 features for classification of ASD vs Comorbid group. Features are named in the same way as in Fig. 8.

[11] Russell A Barkley. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd. *Psychological bulletin*, 121(1):65, 1997.

[12] Thomas J Spencer, Joseph Biederman, and Eric Mick. Attention-deficit/hyperactivity disorder: diagnosis, lifespan, comorbidities, and neurobiology. *Journal of pediatric psychology*, 32(6):631–642, 2007.

[13] Gabrielle Weiss, Lily Hechtman, Thomas Milroy, and Terrye Perlman. Psychiatric status of hyperactives as adults: a controlled prospective 15-year follow-up of 63 hyperactive children. *Journal of the American Academy of Child Psychiatry*, 24(2):211–220, 1985.

[14] Jim Stevenson, Phil Asherson, David Hay, Florence Levy, Jim Swanson, Anita Thapar, and Erik Willcutt. Characterizing the adhd phenotype for genetic studies. *Developmental Science*, 8(2):115–121, 2005.

[15] Genetic and environmental contributions to stability and change of adhd symptoms between 8 and 13 years of age: A longitudinal twin study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(10):1267 – 1275, 2004.

[16] Dsm. diagnostic and statistical manual of mental disorders (dsm-5). `http://www.dsm5.org/`.

[17] Patricia A Rao and Rebecca J Landa. Association between severity of behavioral phenotype and comorbid attention deficit hyperactivity disorder symptoms in children with autism spectrum disorders. *Autism*, page 1362361312470494, 2013.

[18] Qbtest. `https://www.qbtech.com/`.

[19] Peng, Frederick Barrett, Elizabeth Martin, Marina Milonova, Raquel E.

Gur, Ruben C. Gur, Christian Kohler, and Ragini Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods*, 168(1):224 – 238, 2008.

[20] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D.P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Proc. of FG*, Shanghai, China, April 2013. IEEE.

[21] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[22] Antonio Hernndez-Vela, Miguel Reyes, Laura Igual, Josep Moya, Vernica Violant, and Sergio Escalera. Adhd indicators modelling based on dynamic time warping from rgbd data: a feasibility study. In *VI CVC Workshop on the progress of Research & Development, Barcelona, Computer Vision Center*, pages 59–62, 2011.

[23] Marc Parizeau and Rejean Plamondon. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):710–717, 1990.

[24] Ravishankar Sivalingam, Anoop Cherian, Joshua Fasching, Nicholas Walczak, Nathaniel D. Bird, Vassilios Morellas, Barbara Murphy, Kathryn Cullen, Kelvin O. Lim, Guillermo Sapiro, and Nikolaos Papanikolopoulos. A multi-sensor visual tracking system for behavior monitoring of at-risk children. In *ICRA*, pages 1345–1350. IEEE, 2012.

[25] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavioral markers. *CoRR*, abs/1210.7014, 2012.

[26] James M Rehg. Behavior imaging: Using computer vision to study autism. *MVA*, 11:14–21, 2011.

[27] Shyam Sundar Rajagopalan and Roland Goecke. Detecting self-stimulatory behaviours for autism diagnosis. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1470–1474. IEEE, 2014.

[28] Shyam Sundar Rajagopalan, OV Ramana Murthy, Roland Goecke, and Agata Rozga. Play with memeasuring a child's engagement in a social interaction. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[29] Francesca GE Happé. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154, 1994.

[30] Ronald C Kessler, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, EVA Hiripi, Mary J Howes, Robert Jin, Kristina Secnik, Thomas Spencer, et al. The world health organization adult adhd self-report scale (asrs): a short screening scale for use in the general population. *Psychological medicine*, 35(02):245–256, 2005.

[31] Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, 31(1):5–17, 2001.

[32] Ronald C Kessler, Lenard A Adler, Michael J Gruber, Chaitanya A Sarawate, Thomas Spencer, and David L Van Brunt. Validity of the world health organization adult adhd self-report scale (asrs) screener in a representative sample of health plan members. *International journal of methods in psychiatric research*, 16(2):52–65, 2007.

[33] P. Ekman, W. V. Friesen, and J.C. Hager. *Facial action coding system*. Salt Lake City, UT: Research Nexus, 2002.

[34] Igor S. Pandzic and Robert Forchheimer, editors. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc., New York, NY, USA, 2003.