

A Computer Vision based Approach for Understanding Emotional Involvements in Children with Autism Spectrum Disorders

Marco Del Coco, Marco Leo, Pierluigi Carcagnì, Paolo Spagnolo,
Pier Luigi Mazzeo, Massimo Bernava, Flavia Marino, Giovanni Pioggia, Cosimo Distante
National Research Council of Italy,
Institute of Applied Sciences and Intelligent Systems
m.delcoco@isasi.cnr.it

Abstract

It has been proved that Autism Spectrum Disorders (ASD) are associated with amplified emotional responses and poor emotional control. Underlying mechanisms and characteristics of these difficulties in using, sharing and responding to emotions are still not understood. This is because advanced computational approaches for studying details of facial expressions have been based on the use of invasive instruments (such markers for motion capture or Electromyographs) that can affect the behaviors and, above all, restrict the possibility to implement diagnostic and evaluation tools. Recent non-invasive technological frameworks based on computer vision can be applied to overcome this knowledge gap and this paper is right aimed at demonstrating how facial measurements from images can be exploited to compare how ASD children react to external stimuli with respect a control set of children. This paper has a double layer of contribution: on the one hand it aims at proposing the use of a single-camera system for facial expression analysis and, on the other hand, it presents a study on how extracted facial data could be used to analyze how the overall and local facial dynamics of children with ASD differ from their typically developing peers. In other words, this study explores the feasibility of the introduction of numerical approaches for the diagnosis and evaluation of autistic spectrum disorders in preschool children.

1. Introduction

Current diagnostic criteria (e.g., ICD-10 and DSM-IV) list marked impairments in the use of facial expression in social interaction as evidences of Autistic Spectrum Disorders (ASD). It has been in fact proved that ASD are associated with amplified emotional responses and poor emotional control but underlying mechanisms and characteristics of these difficulties in using, sharing and responding to

emotion are still not understood. Emotion regulation (ER) strategies can be used to understand emotional problems in ASD. ER is a term generally used to describe a person's ability to effectively manage and respond to an emotional experience, and then ER strategies define automatic or intentional modifications of a person's emotional state that promotes adaptive or goal-directed behavior [25]. ER strategies utilize stimuli that resemble real-life situations in order to elicit real-time emotional activation that provide quantitative and qualitative assessments of individual differences in emotional reactivity and regulation [32]. Small humanoid robot with simplified human-like features have been also used to stimulate the interactions with children [2]. Methods to study ER are based either on naturalistic observation of facial and vocal indices or on clinical measurements e.g., Heart rate variability, respiratory sinus arrhythmia, and functional magnetic resonance imaging. A systematic literature review of Emotion Regulation (ER) measurement in individuals with autism spectrum disorder has been proposed in [33] to identify the various ways and processes of ER that have been studied in individuals with ASD. The paper highlights the main limitations to assess ER in individuals with ASD: on the one hand, the reliability of clinical measurements is still under debate whereas, on the other hand, methods based on observations are affected by difficulty to interpret emotions without defining the context of the child's baseline behaviors and emotional expressions. A way to overcome these drawbacks is to get measurements of appearance cues related to emotions [21]. This is the motivation of recently computational studies based on motion capture data that have been carried out bringing to the observations that High Functioning Autism children have reduced complexity in the dynamic facial behavior, arising primarily from the eye region [15]. The study made use of thirty two reflective markers that were affixed to the face of each participant; the movement of these markers was recorded by six infrared motion capture cameras at 100 frames per

second. A few studies [30, 23] involved quantitative methods to analyze the facial expression using electrophysiological sensors like electromyography (EMG). Unfortunately, markers and sensors are intrusively placed on the facial skin and may potentially inhibit spontaneous facial expressions.

Recent advances in computer vision and machine learning brought to more and more affordable solutions for facial analysis [8, 5] that paved the way for developing non-invasive technological frameworks which can be applied to extract facial measurements in a non-invasive way. This is a pioneering research area and a very few works can be found on related topics. In [31] two non-intrusive optical imaging sensors, e.g., a video camera and a 3D optical camera, have been employed during a pilot study to capture 2D and 3D facial images of participants in response to visual stimuli, respectively. An expression training interface which evaluates the imitation of facial expressions and head movements has been proposed in [1] and in [12].

This paper is aimed at demonstrating how recent computer vision frameworks can be exploited to compare how ASD children react to external stimuli with respect to a control set of children.

This paper has a double layer of contribution: on the one hand it aims at proposing the use of a single-camera system for facial expression analysis and, on the other hand, it presents a study on how extracted facial data could be used to analyze how the overall and local facial dynamics of children with ASD differ from their typically developing peers. In other words, this study explores the feasibility of the introduction of numerical approaches for the diagnosis and evaluation of autistic spectrum disorders in preschool children.

To do that, in the paper two small groups of children (affected by ASD and Typically Developed) were acquired from a web-cam while watching cartoons properly chosen to elicit three emotions: happiness, fear and sadness. In the first experimental phase the facial behavioral complexity was analyzed in order to point out global and local differences between groups in the emotional reactions, without possible biases introduced by the wearable and/or invasive acquisition tools used so far.

In the second experimental phase an emotional similarity score was computed. This might be of interest to further investigate the possibility to determine an objective metric that can automatically distinguish children belonging to the two different groups and eventually also give a diagnosis or assessment score.

The rest of the paper is organized as follows: Section 2 describes computer vision components and computational strategies involved in the analysis of children's faces, Section 3 reports experimental setup whereas in Section 4 experimental outcomes are detailed and discussed. Section 5 concludes the paper.

2. Computer Vision Module

Computer Vision module is made up by four main components aiming at face detection, facial landmark detection, multi-face tracking and Facial Action Unit extraction. The employed framework, inspired by the algorithmic procedure proposed in [5], starts with a *face detection* step. If a face is detected it is subsequently analyzed by a *Facial Landmark Detection* block. In order to deal with the case in which more than one person is simultaneously in the scene, the facial models assigned to any singular face in the scene are exploited for *Multi-face tracking*. As a last step, the detected face and landmarks are processed in order to compute the Facial Action Unit intensities.

2.1. Face Detection

The face detection is performed by making use of Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid and a sliding window detection scheme. This type of object detector is fairly general and capable of detecting semi-rigid objects as the human faces are [18].

2.2. Facial Landmark Detection

Once a face is detected, facial landmarks are detected by Conditional Local Neural Field (CLNF) as proposed in [4]. CLNF is an instance of the Constrained Local Models (CLM) proposed in [9] and consists of two main components:

- a Point Distribution Model (PDM) aimed to capture landmarks shape variations;
- patch experts, improved respect to CLM, in order to capture appearance variations of each landmark and suitable for *in-the-wild* scenarios.

A CLM model can be described by a set of parameters $\mathbf{p} = [s, \mathbf{R}, \mathbf{q}, \mathbf{t}]$ that can be varied in order to acquire various instances of the model: the scale factor s ; object rotation \mathbf{R} (first two rows of a 3D rotation matrix); 2D translation \mathbf{t} ; a vector describing non-rigid variation of shape \mathbf{q} . The point distribution model (PDM) is: $\mathbf{x}_i = s \cdot \mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i \mathbf{q}) + \mathbf{t}$. Here $\mathbf{x}_i = (x, y)$ denotes the 2D location of the i^{th} feature point in an image, $\bar{\mathbf{x}}_i = (X, Y, Z)$ is the mean value of the i^{th} element of the PDM in the 3D reference frame, and the vector Φ_i is the i^{th} eigenvector obtained from the training set that describes the linear variations of non-rigid shape of this feature point.

In CLM (and CLNF), the maximum a posterior probability (MAP) of the face model parameters \mathbf{p} given an initial location of the parameters determined by a face detection step is estimated.

The solution in use improves the standard approach by means of the training of separate sets of point distributions

and patch expert models for eyes, lips and eyebrows. As a successive step it fits the landmarks detected with individual models to a joint PDM. The tracking phase is supported by a face validation step aimed to avoid face leading or face drifting over long period of time. To this end the system employs a Convolutional Neural Network (CNN) that, given a face aligned using a piece-wise affine warp, predicts the expected landmark detection error. In this way the models can be reset when the validation step fails. As a final enforcement, a multiple initialization hypotheses (at different orientations) is employed in order to pick the best converge likelihood and manage challenging in the wild acquired images.

The used PDM (36 non-rigid and 6 rigid shape parameters) and CNN are both trained on the LFPW [6] and Helen [19] training sets. On the other hand, the CLNF patch experts are trained on Multi-PIE [14], LFPW [6] and Helen [19] training sets. A key point for the robustness of the proposed approach is the use of 28 sets of patch experts trained at different scales and views that allow to handle different images resolution of the face under analysis as well as head rotations and consequent self occlusions.

2.3. Multi-Face Tracker

In order to deal with the case in which more than one person is simultaneously in the scene, the CLNF models assigned to any singular face in the scene are exploited for Multi-face tracking. This is achieved by performing, every a certain number of frames, a new face detection procedure and then by checking if detected bounding boxes overlap the faces already tracked. For the boxes that do not overlap any predefined bounding box, a new CLNF model is created and used to track the new detected faces. When a new CLNF model is instantiated, besides information related to landmarks, bounding boxes and pose, an unique identifier (UI) for indexing the face tracked is used. CLNF models, one for each single face, are updated in a parallel fashion in order to perform multi-face tracking. In case of extreme pose of the tracked face, CLNF approach is unable to detect the facial landmarks. In this case, tracking is no longer performed and the tracked face gets out from the tracking process. In order to be tracked again, we have to wait when a face detection takes again. In this step, if the face results in a near frontal face pose it is the detected and tracking starts again but assigning a new UI. In order to assign the same UI, a re-identification process is needed. To such purpose, a Deep Convolution Neural Network approach has been employed. In particular, assumed that throughout the video the same subjects are present and that a variation of the UIs is due to extreme pose changes, every face is processed frame by frame by means of the VGG-Face CNN [27] where, in order to extract a robust descriptor to be used for the re-identification task, the last interconnect layer

(FC7) is used as features vector. Hence, face descriptors of the current frame are compared with the previous frames' ones by means of minimal Euclidian distance. The smallest distance is related to descriptors of the same face.

2.4. Action Unit Detection

The reliability of an action unit classifier depends largely on the employed training data and its ability to estimate facial expressions of a subject when his neutral one is unknown. The proposed solution exploits the idea proposed in [3] where the authors introduce a real-time Facial Action Unit intensity estimation and occurrence detection system based on geometry features (shape parameters and landmark locations computed by the CLNF) and appearance (Histograms of Oriented Gradients). The first step for a correct detection of a AU presence and intensity is the mapping of the detected face to a common reference frame. To this end the currently detected landmarks are transformed to a representation of frontal landmark from a natural expression (a projection of mean shape from a 3D PDM). The resulting is a 112×112 pixel image of the face with 45 pixel inter-pupillary distance. In order to remove non-facial information from the image, a masking of the image is performed using a convex hull surrounding the aligned feature points. The aligned face results in a 112×112 image ready for appearance features extraction. In this step, Histograms of Oriented Gradients (HOGs) are extracted as proposed in [10]. Blocks of 2×2 cells, of 8×8 pixels are employed and lead to 12×12 blocks of 31 dimensional histograms. The final vector size is of 4464 elements describing the face subsequently reduced to 1391 elements by means of a Principal Component Analysis (PCA) approach. The non-rigid shape parameters and landmark locations in object space inferred during CLNF model tracking are used as geometry based features that results in a 227 dimensional vector describing geometry. The complete features vector is then made up by the concatenation of the geometry and appearance ones. In order to account for personal differences the median value of the features (observed so far in online case and overall for offline processing) is subtracted from the estimates in the current frame. The last step for the AU detection and intensity estimation is obtained, respectively, with a Support Vector Machines (SVM) and Support Vector Regression (SVR). In both cases, linear kernels are employed. The models used in the proposed approach are trained on DISFA [24], SEMAINE [26] and BP4D [34] datasets. Where the AU labels overlap across multiple datasets we train on them jointly.

Inspired by the observation that only a few facial parts are active in expression disclosure (e.g. around mouth, eye), previous works discovered the common and specific patches which are important to discriminate all the expressions and only a particular expression, respectively [35][11]. In light

of this, in this study the 18 action units that incorporate the most significant variations of eye brows, eye lids, cheeks and lips ensuring the ability to see the expressions of the main emotional states are computed. The list of recognized AUs is presented in Table 1.

Moreover, existing AU predictors tend to under- or over-estimate AU values for specific person. To avoid this prediction errors, the lowest n_{th} percentile (learned on validation data) of the predictions on a specific person has been subtracted from all of the predictions.

2.5. Data Analysis Module

To objectively quantify AU signaling over time, the Shannon entropy, which measures the complexity (i.e., average uncertainty) of a signal is used. The Shannon entropy is calculated for each AU across time as suggested in [16]. Each AU is treated as a random variable X that takes on only finitely many values and then its Shannon entropy is defined by the formula

$$H(X) = - \sum_i p(x_i) \log(p(x_i))$$

with the convention that $0 \log \frac{1}{0} = 0$.

The $p(x_i)$ is the probability to have x_i in the stream of observed values and it is computed by an initial discretization of observed action unit values in N bins and then by estimating the associated Probability Density Function. This is done for each AU and child in each time interval related to the same stimulus.

This way the complexity of the child's reaction to the given stimulus can be computed.



















Behavioral Similarity among ASD and TD group is estimated by using an initial alignment by Dynamic Time Warping [17].

Let X^N be the set of discrete-time time series taking values in an arbitrary space X . Taking two time series $u = (u_1, \dots, u_n)$ and $v = (v_1, \dots, v_m)$ of lengths n and m respectively, an alignment π has length p and $pn + m1$ since the two series have $n + m$ points and they are matched at least at one point of time. An alignment π is a pair of increasing integral vectors (π_1, π_2) of length p such that $1 = \pi_1(1) \leq \dots \leq \pi_1(p) = n$ and $1 = \pi_2(1) \leq \dots \leq \pi_2(p) = m$, with unitary increments and no simultaneous repetitions. Coordinates of π are also known as warping functions.

Now, let $|\pi|$ denote the length of alignment π . The cost can be defined by means of a local divergence that measures the discrepancy between any two points u_i and v_j of vectors u and v :

$$D_{u,v}(\pi) = \sum_i^{|\pi|} \phi(u_{\pi_1(i)}, v_{\pi_2(i)})$$

Table 1. Action Units computed by the framework

| AU | Full Name | Example |
|------|----------------------|---|
| AU1 | Inner brow raiser |  |
| AU2 | Outer brow raiser |  |
| AU4 | Brow lowerer |  |
| AU5 | Upper lid raiser |  |
| AU6 | Cheek raiser |  |
| AU7 | Lid tightener |  |
| AU9 | Nose wrinkler |  |
| AU10 | Upper lip raiser |  |
| AU12 | Lip corner puller |  |
| AU14 | Dimpler |  |
| AU15 | Lip corner depressor |  |
| AU17 | Chin raiser |  |
| AU20 | Lip stretched |  |
| AU23 | Lip tightener |  |
| AU25 | Lips part |  |
| AU26 | Jaw drop |  |
| AU28 | Lip suck |  |
| AU45 | Blink |  |

were $\phi(x, y)$ = The Global Alignment (GA) kernel is defined as the sum of exponentiated and sign changed costs of the individual alignments:

$$k_{GA}(u, v) = \sum e^{D_{u,v}(\pi)}$$

were the sum takes over the set of all alignments between two time series of length n and m . In the following the local kernel e^{ϕ_σ} is used where

$$\phi_\sigma(x, y) = \frac{1}{2\sigma^2} \|x - y\|^2 + \log(2 - e^{-\frac{\|x-y\|^2}{2\sigma^2}})$$

Finally a distance metric is obtained from the above kernel by using the standard transformation described in the following equation:

$$d(u, v) = k_{GA}(u, u) + k_{GA}(v, v) - 2k_{GA}(u, v)$$

3. Experimental Setup

Five children with ASD, aged 48-65 months (average 65.38, standard deviation 15.86), were enrolled in the study. ASD children were tested at the clinical facilities within the National Research Council of Italy (CNR), Messina, Italy. ASD diagnosis was made according to the DSM-5 criteria 1 [29] by an experienced multidisciplinary team including two child psychiatrists and 2 developmental psychologists. The Autism Diagnostic Observation Schedule - Second Edition (ADOS-2) [22] was used as part of the diagnostic assessment. The Griffiths Mental Development Scale (GMDS) was used to assess the Developmental Quotient (DQ) [13]. Developmental Quotient for the involved ASD children was 92, 78, 71, 68 and 42 respectively. The Typically Developed (TD) control group comprised five children with age and gender corresponding to the above mentioned ASD group. Parents (or Guardian) signed an informed consent form for agreeing with the children participation in this research study.

Each children (TD and ASD) was asked to watch, together with parents and therapists, a sequence of 9 videos taken from famous cartoons. The sequence alternates videos eliciting emotions of happiness, fear and sadness. Videos were supplied in a Lab able to simulate a child home environment while embodying disappearing technology to ecologically quantify physiological and behavioural variables, coach parents and apply personalized treatment. The videos ran on a monitor TV and a web cam positioned on top of the monitor was used to acquire faces of persons watching. The duration of each video is a priori known and this makes possible to directly label the acquired data with the corresponding ideal emotional state that should have been elicited by each video.

4. Experimental Results

The first experimental phase was aimed at analyzing the complexity of emotional reaction to external stimuli for

Table 2. Computed entropy score (global and local scores)

| | Happiness | | Fear | | Sadness | |
|------------|-----------|------|------|------|---------|------|
| | ASD | TD | ASD | TD | ASD | TD |
| Upper face | 1572 | 2070 | 1409 | 1690 | 1481 | 1767 |
| Lower face | 1889 | 2204 | 1665 | 1887 | 1743 | 1861 |
| Overall | 1776 | 2156 | 1574 | 1819 | 1644 | 1827 |

both ASD and TD groups. To do that each of the facial action units extracted as described in subsection 2.4 was handled as a dynamic mono dimensional signal and its entropy H was computed as described in Section 2.5.

The overall facial complexity was computed by averaging all the obtained entropy values whereas two local entropy scores were computed by averaging the entropy computed action unit belonging to the upper (eye region) and lower (mouth region) parts of the face. This choice was made according to the clinical evidence that the considered emotions are usually expressed by changing eyes and mouth configuration [28].

Computed entropy scores are reported in Table 2: columns report, for each elicited emotion, the entropy scores for ASD and TD groups whereas rows refer to the considered facial region or to the global (overall) facial entropy. The entropy values computed on the overall action units are very interesting. It is in fact evident that, independently from the elicited emotion, TD children exhibited more facial behavioral complexity. In particular eye and mouth regions contributed to this differential whereas cheek dynamics was comparable (fear) or even higher (happiness and sadness) in the ASD group. These results are broadly consistent with those expected taking into account both clinical studies or experimental evidences derived through invasive methods for data acquisition [7]. This way the outcomes of the proposed non-invasive approach pose an interesting perspective to make possible the analysis of facial dynamics by using just computer vision based algorithms.

In the second experimental phase the similarity between emotional behaviors of children under investigation was computed by the approach based on Dynamic Time Warping described in Section 2.5.

Numerical results of this experimental phase are in table 3. In particular the computed values derived by using the introduced similarity metric are reported. Each value corresponds to the average of the ones relative to an elicited emotion (indicated in the first row in the heading) for couples of children selected in the given groups (pointed out in the second row of the heading) and considered for upper, lower and overall face part (as described in the leftmost column).

It is evident that the introduced metric highlights that

facial behaviors are more similar when children belong to the TD group than in the case in which the children belong to different groups. This is true for both upper and lower face part and, of course, it is even more emphasized when the overall face is taken under consideration. It is also interesting to observe as lower face seems more significant than upper face to distinguish between TD and ASD children. This means that, through deeper experimental investigations, the introduced approach could bring to affordable non-invasive measurement for the diagnosis and assessment of autism spectrum disorders.

5. Conclusions

In the paper two small groups of children (affected by ASD and Typically Developed) were acquired from a webcam while watching cartoons properly chosen to elicit three emotions: happiness, fear and sadness. The main aim of the paper was to demonstrate if computer vision based approaches for facial feature analysis could help to understand emotional behaviors in children with the interesting perspective of introducing a computational approach for diagnosis and assessment of autism spectrum disorders. Two experimental phases were carried out. In the first experimental phase the facial behavioral complexity was analyzed in order to point out global and local differences between groups in the emotional reactions, without possible biases introduced by the wearable and/or invasive acquisition tools used so far. In the second experimental phase an emotional similarity score was computed trying to find out if it might be of interest to further investigate to determine an objective metric that can automatically distinguish children belonging to the two different groups and eventually also give a diagnosis or assessment score. Future works will deal with a more systematic analysis of data by further campaigns of data acquisition comprising a more numerous set of ASD and TD. Also additional facial cues such as gaze will be considered [20]. Moreover, parent-child interaction will be analyzed.

References

- [1] A. Adams and P. Robinson. Expression training for complex emotions using facial expressions and head movements. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 784–786. IEEE, 2015. 2
- [2] S. M. Anzalone, S. Boucenna, D. Cohen, M. Chetouani, et al. Autism assessment through a small humanoid robot. In *Proc. HRI: A Bridge between Robotics and Neuroscience, Workshop of the 9th ACM/IEEE Int. Conf. Human-robot Interaction*, pages 1–2, 2014. 1
- [3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015. 3
- [4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013. 2
- [5] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 2
- [6] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. 3
- [7] D. Bone, T. Chaspari, and S. Narayanan. Behavioral signal processing and autism. *Autism Imaging and Devices*, page 319, 2017. 5
- [8] P. Carcagnì, M. Coco, M. Leo, and C. Distantè. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1):1, 2015. 2
- [9] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 1, page 3, 2006. 2
- [10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 3
- [11] M. Ghayoumi and A. K. Bansal. Unifying geometric features and facial action units for improved performance of facial expression analysis. *arXiv preprint arXiv:1606.00822*, 2016. 3
- [12] I. Gordon, M. D. Pierce, M. S. Bartlett, and J. W. Tanaka. Training facial expression production in children on the autism spectrum. *Journal of Autism and Developmental Disorders*, 44(10):2486–2498, 2014. 2
- [13] R. Griffith, D. Luiz, A. for Research in Infant, and C. Development. *Griffiths Mental Development Scales, Extended Revised: GMDS-ER; Two to Eight Years*. Hogrefe, the Test People, 2006. 5
- [14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3
- [15] T. Guha, Z. Yang, R. B. Grossman, and S. S. Narayanan. A computational study of expressive facial dynamics in children with autism. *IEEE Transactions on Affective Computing*, 2017. 1
- [16] R. E. Jack, O. G. Garrod, and P. G. Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014. 4
- [17] L. A. Jeni, A. Lőrincz, Z. Szabó, J. F. Cohn, and T. Kanade. Spatio-temporal event classification using time-series kernel based structured sparsity. In *Computer vision-ECCV - European Conference on Computer Vision: proceedings. European Conference on Computer Vision*, volume 2014, page 135. NIH Public Access, 2014. 4
- [18] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. 2

Table 3. Computed similarity metric

| | Happiness | | Fear | | Sadness | |
|------------|------------------|-------|-------------|-------|----------------|-------|
| | ASD/TD | TD/TD | ASD/TD | TD/TD | ASD/TD | TD/TD |
| Upper face | 2121 | 2013 | 2178 | 2028 | 2094 | 1848 |
| Lower face | 3071 | 2825 | 2693 | 2459 | 4207 | 2023 |
| Overall | 4919 | 4331 | 4589 | 4371 | 4304 | 3718 |

- [19] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012. 3
- [20] M. Leo, D. Cazzato, T. De Marco, and C. Distanto. Unsupervised eye pupil localization through differential geometry and local self-similarity matching. *PloS one*, 9(8):e102829, 2014. 6
- [21] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1 – 15, 2017. 1
- [22] C. Lord, M. Rutter, P. DiLavore, S. Risi, R. Luyster, K. Gotham, S. Bishop, and G. W. *ADOS-2: Autism Diagnostic Observation Schedule : Manual*. 2012. 5
- [23] D. Mathersul, S. McDonald, and J. A. Rushby. Automatic facial responses to affective stimuli in high-functioning adults with autism spectrum disorder. *Physiology & behavior*, 109:14–22, 2013. 2
- [24] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3
- [25] C. A. Mazefsky, J. Herrington, M. Siegel, A. Scarpa, B. B. Maddox, L. Scahill, and S. W. White. The role of emotion regulation in autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 52(7):679–688, 2013. 1
- [26] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010. 3
- [27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 3
- [28] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, and J. Piven. Visual scanning of faces in autism. *Journal of autism and developmental disorders*, 32(4):249–261, 2002. 5
- [29] D. A. Regier, E. A. Kuhl, and D. J. Kupfer. The dsm-5: Classification and criteria changes. *World Psychiatry*, 12(2):92–98, 2013. 5
- [30] A. Rozga, T. Z. King, R. W. Vuduc, and D. L. Robins. Undifferentiated facial electromyography responses to dynamic, audio-visual emotion displays in individuals with autism spectrum disorders. *Developmental science*, 16(4):499–514, 2013. 2
- [31] M. D. Samad, J. L. Bobzien, J. W. Harrington, and K. M. Iftekharuddin. Non-intrusive optical imaging of face to probe physiological traits in autism spectrum disorder. *Optics Laser Technology*, 77:221 – 228, 2016. 2
- [32] A. C. Samson, A. Y. Hardan, R. W. Podell, J. M. Phillips, and J. J. Gross. Emotion regulation in children and adolescents with autism spectrum disorder. *Autism Research*, 8(1):9–18, 2015. 1
- [33] J. A. Weiss, K. Thomson, and L. Chan. A systematic literature review of emotion regulation measurement in individuals with autism spectrum disorder. *Autism Research*, 7(6):629–648, 2014. 1
- [34] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 3
- [35] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE, 2012. 3