CLASSIFYING AUTISM SPECTRUM DISORDER BASED ON SCANPATHS AND SALIENCY

Mikhail Startsev, Michael Dorr

Technical University of Munich, Germany

ABSTRACT

Individuals suffering from autism spectrum disorder (ASD) demonstrate viewing patterns that are often different from those exhibited by control subjects, especially in the context of emotional or social stimuli. Previous studies with such content typically relied on precise hand-labelling of the regions of interest in displayed images, thus limiting the transferability of performed analyses onto new, unlabelled images. In contrast to this, we propose an approach that classifies the viewing behaviour of individuals as likely associated with either ASD or typical development in a fully automatic fashion, relying on scanpath features and analytically predicted saliency. Our analysis further demonstrates that gaze data for images with multiple faces have a substantially higher discriminative power than other image groups.

Index Terms— Autism spectrum disorder, image viewing, saliency, face viewing, eye tracking

1. INTRODUCTION

Autism spectrum disorder (ASD) is a range of neurodevelopmental disorders including several distinct conditions that vary in severity. ASD is associated with limited social and communication skills, such as being unresponsive to the individual's name being called or other social stimuli [1], reduced tendency to gaze-following [2], generally looking less at faces in social scenes [3, 4] and eyes in particular [5, 6], or difficulties identifying emotions [5, 6] and even recognising faces in general [6]. The findings are not always consistent between studies and set-ups, however: In some cases no significant differences between the ASD and typically developing (TD) groups in terms of attention allocation to faces in general [2, 6] or eyes in particular [2] emerge.

One disadvantage of analysing the eye tracking recordings in terms of specific regions of interest lies in the necessity to annotate these in the images. The authors of [3] manually labelled body, face, eyes, and mouth areas of the images. In [2], each image was subdivided in up to six regions, including the non-trivial labelling of the depicted person's approximate gaze direction and object of interest in the scene. In [4] the images were manually segmented into objects, the semantic properties of which also had to be annotated. In [1] two assistants monitored experiment participants and recorded video footage. With modern techniques for facial landmark detection, however, the need for laborious manual image processing could be alleviated. Part of this work's contribution consists, therefore, in the application of automatic face detection when analysing scanpaths of ASD and TD individuals, suggesting that automated analyses of social stimuli could replace manual annotations.

Previous work [2, 7] studied the effect of low-level saliency in participants with ASD and schizophrenia, respectively, with [2] finding no difference to controls. [7] found significantly reduced saliency at gaze locations in the schizophrenia group for video but not picture viewing. [4] compared different-level image features on pixels being empirically salient in ASD or TD groups, finding that the ASD group was affected by centre bias and low-level saliency more than TD. In this study, we investigated both low- and high-level saliency maps, analysing their importance for distinguishing the ASD and TD groups (similarly to [4]).

Attention allocation (relative to some pre-defined areas of interest or saliency), however, is not the only difference studied in the context of neurological conditions. An additional level of gaze behaviour analysis is enabled by extracting the eye movements that constitute the gaze trace. The characteristics of these "events" can also be examined to differentiate between the populations. In [2] basic fixations and saccade statistics were compared between the groups (without a significant effect observed for adolescent participants). In principle, dynamic eye movements can help identify individuals with certain vision- or perception-related disorders. Smooth pursuit (following a moving object with the eyes) is, for example, typically impaired in schizophrenia patients [7, 8].

This work, therefore, combines features from various domains for the purpose of automatically distinguishing the scanpaths belonging to subjects with ASD from those of typically developing controls. We used eye movement statistics, saliency-based, as well as face-based features. The separability of the two classes achieved by our model is relatively high (cross-validated 75% AUC), considering that this classification is based on a viewing of a single arbitrary image. In a real scenario, a series of images would typically be evaluated, which should substantially increase the classification quality.

Using our model, we also observed that the viewing patterns of the ASD and TD groups for images with multiple faces allow for significantly better classification power of the

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00122

model. This agrees with the prior findings in the literature and points out the benefits of automatically pre-selecting image subsets that could be used to better support diagnosis or before applying classification techniques.

2. METHODS

2.1. Data set and evaluation

The data set used in this work was collected in the context of an ICME 2019 Grand Challenge "Saliency4ASD: Visual attention modeling for Autism Spectrum Disorder" (full details can be found in [9]). The eye tracking was performed with a Tobii T120 tracker, fixations detected with Tobii Pro Studio. The resulting scanpath data were available for analysis: Each recording was represented by a sequence of fixations, each described by their x, y location on the image and duration.

Both ASD and TD groups included 14 subjects each (both groups with the mean age of 8), all viewing 300 images (3 s display duration) in the training set. The stimuli represent a subset of the MIT1003 data set [10], which was originally developed for image saliency research, and contains a diverse set of naturalistic targets (people, animals, indoor and outdoor objects) that should be representative of everyday life. Some of the subjects' recordings were discarded due to lack of attention, but overall the training set is fairly well-balanced (3761 and 3837 scanpaths for ASD and TD, respectively).

During model development on the training set, we used bootstrapping to fairly assess the quality of the classifier: We repeatedly (100 times) sampled a small balanced test set of scanpaths (250 from each group), while the others were used for training. For each test set, we constructed the corresponding receiver operating characteristic (ROC; using the model's score – assigned probability – for the ASD class) and computed its area under the curve (AUC). We then report the AUC score mean and 95% confidence interval (over all iterations).

Since the Grand Challenge test data set would contain images that were not present in the training set, we wanted to assess our model's ability to generalise to unseen stimuli and not just unseen scanpaths. For this purpose, we performed the bootstrapping by sampling not individual scanpaths, but all of the scanpaths corresponding to some images at once. To achieve at least 250 scanpaths per group, on average twenty images would have to be considered as the test set on each iteration. We randomly selected exactly 250 scanpaths per group in order to give equal weight to the scores of each run.

2.2. Features for classification

This section describes the features we considered during the development of our model. Not all of these were selected for the final feature set, but we make important observations based on them as well. We note that in the set-up of the challenge, each scanpath for each image was to be treated individually and the information about the scanpaths' correspondence to unique individuals was not provided. Therefore, the features could only rely on (i) the recorded sequence of fixations and (ii) the stimulus image for this gaze trace.

2.2.1. Scanpath features

First, we computed the basic fixation statistics of each scanpath: their number, total and average duration, and average distance from the centre of the corresponding image or from the centre of the scanpath (mean of all scanpath points). For the transitions between subsequent fixations, we computed the total length of the scanpath and the average amplitude (i.e. approximate saccade amplitude statistics, even though no explicit saccade information was provided).

2.2.2. Saliency features

Similar to [7, 2], we tested whether the gaze of the two groups is differently driven by stimulus saliency. Those works used comparatively low-level saliency maps, and [2] did not find group differences between ASD and TD. Here, we tested two low-level bottom-up saliency models as well (GBVS [11] and Itti-Koch [12]), and found that the group differences for a high-level saliency model – the 2018 version of the [13] SAM-ResNet – were much higher than for the low-level models: For example, the AUC for the mean fixation saliency value for SAM-ResNet was 58% vs. 53% and 55% for the [12] and [11] models, respectively. Furthermore, keeping the low-level saliency-based features did not improve classification over what can be achieved with just high-level saliency.

Therefore, our final model relies only on the saliency features computed based on the [13] predictions, blurred with a Gaussian kernel ($\sigma = 43$ px, ca. 1° of the visual field). For this saliency map, we computed the mean (similar to [2]), maximal, and sum of the saliency values of each scanpath, the mean and total sum of fixation durations weighted by the corresponding saliency values, the saliency value of the first fixation (as in [2]), as well as the index of the first fixation to have reached 75 and 90% of the largest saliency value on the viewed image (or 20, if no fixation reaches the corresponding values). The latter features should describe the subjects' prioritisation of the areas of interest. We also created empirical saliency maps for each subject and compared those to the prediction of SAM-ResNet with typical measures [14]: normalised scanpath saliency and Kullback-Leibler divergence.

2.2.3. Face-driven features

As ASD is often investigated in the context of face or social scene viewing [15, 2, 3], we incorporated face gazing features in our model. To avoid manual annotation, a convolutional neural network-based face detector (part of `http://dlib.net`) was employed. Its detected bounding boxes were enlarged by 50% along each axis to compensate for potential gaze tracking noise and face detection imprecision.

We computed the number and share of fixations on faces and the total time (and total fixation time share) spent looking at faces, as well as their normalised variants – divided by the share of the corresponding image that was covered by face rectangles (not accounting for potential overlap). The normalisation procedure is similar to that in [2]. If multiple faces were detected, the entropy of the distribution of attention towards the automatically detected faces (as number and duration of fixations, also with face rectangle area-based normalisation) was computed. For images where some of these statistics were inapplicable, a value of -1 was used instead.

2.3. Classification

Using the Scikit-learn library (`http://scikit-learn.org/`), we trained a random forest to classify individual scanpaths (using different feature combinations) as belonging to a TD or an ASD subject. We empirically determined that the performance of our model was not strongly influenced by the number of the trees in the ensemble (1000 in the final model), but benefited from limiting the tree depth (for the final model: at least 5 samples in each leaf and a maximal depth of 10). We did not use sequence modelling for scanpath classification (as e.g. [16]) as the data set we used consists of fixation sequences (often 1 or 2 points only), and not all gaze samples.

3. RESULTS

3.1. General observations

Even though during image viewing fixations and saccades should almost fully account for the viewing time, and saccades would normally occur with a similar frequency for all observers, we noticed that the total time of fixations was a very strong feature (66% AUC for this one statistic; shorter in the ASD group). Depending on the fixation detector, this feature would actually be affected by a rather comprehensive set of factors: saccade and blink frequency, saccade amplitude distribution, noise level in the recording, etc. Other studies [2] computed similar measures, but found no statistical differences between the groups.

While previous studies found reduced exploratory behaviour in individuals with ASD [15], most of our scene exploration features pointed towards the inverse: The ASD group fixated further away from the centre of both the image and own scanpath, making larger transitions between fixations. This is in line with the findings in [17], where ASD subjects explored more elements of the web pages. The total length of the scanpath was lower in the ASD group in our data (same as found in [15]), but this can be attributed to a lower number of recorded fixations in this group in our data.

Both face gazing and multi-face attention entropy features improved the classification quality when added to scanpath characteristics, indicating the usefulness of such automated analyses and the presence of group differences in terms of spontaneous face gazing and scene exploration with multiple faces. They did not, however, deliver an improvement in the presence of saliency features, and were therefore excluded from the final model to reduce its complexity.

With more data available or in richer stimulus domains other eye movement features may be investigated. E.g. several prior works [18, 19] have investigated smooth pursuit behaviour of children and young adults with ASD, but findings can differ between artificial and naturalistic stimuli [7].

3.2. Classification performance

Tested via bootstrapping, our model achieved a cross-validated 75% AUC (all subjects' scanpaths for all images ranked by the predicted probability of the ASD class) on the training set (95%-CI – confidence interval – $[0.71; 0.78]$) Due to the fact that the test data set contained stimuli from different image data sets [9], the features our model works with were significantly different compared to the training data (according to the two-sample Kolmogorov-Smirnov test), our model performance was lower on the test set of the Grand Challenge: 63.9% AUC at 63.5% F1 score.

To further investigate the importance of the image content, we separately tested the classification performance on the image subsets (of the training set) that contained zero, one, or at least two faces (detected as in Section 2.2.3). We found that our model's performance on the subset of images containing multiple faces was significantly better than for other subsets or for the whole data set (76.9% AUC on average, 95%-CI $[0.74; 0.81]$ Mann-Whitney U-test $p < 10^{-7}$); notably, group differentiation analysis based on single-face images was significantly poorer than for other conditions (71.6% AUC on average, 95%-CI $[0.69; 0.77]$, $p < 10^{-2}$). Interestingly, when examining feature importances of our model, all high-level saliency features were much more useful on this subset than on others (e.g. saliency features used in the trained random forest classifiers on average 20% more frequently than for the full image set and 28% more than for images without faces).

The source code for the full pipeline, as well as the model trained on the training set of the challenge and the individual feature importances, are available at `https://github.com/MikhailStartsev/ASD_classification/`.

4. CONCLUSIONS

In this work, we have developed, tested, and made publicly available a fully automatic system that distinguishes children with autism spectrum disorder from the typically developing controls based on the scanpaths they make for arbitrary image viewing, thus helping to pave the way to unobtrusive fully automated ambulatory monitoring of the symptoms. To this end, we examined the gaze data collected during image free-viewing and employed scanpath features as well as those based on (automatic) face detection and saliency prediction.

Our model and its analysis allow for meaningfully choosing images that should be presented to the patients to be able to better support diagnosis, as well as ranking the features by their contribution to classification quality, thus revealing the points of difference between the ASD and TD groups that are of particular importance to their correct classification.

5. ACKNOWLEDGEMENTS

This research was supported by the Elite Network Bavaria.

6. REFERENCES

[1] Geraldine Dawson, Andrew N. Meltzoff, Julie Osterling, Julie Rinaldi, and Emily Brown, "Children with autism fail to orient to naturally occurring social stimuli," *Journal of Autism and Developmental Disorders*, vol. 28, no. 6, pp. 479–485, Dec 1998.

[2] S. Fletcher-Watson, S.R. Leekam, V. Benson, M.C. Frank, and J.M. Findlay, "Eye-movements reveal attention to social information in autism spectrum disorder," *Neuropsychologia*, vol. 47, no. 1, pp. 248 – 257, 2009.

[3] Deborah M. Riby and Peter J.B. Hancock, "Viewing it differently: Social scene perception in Williams syndrome and autism," *Neuropsychologia*, vol. 46, no. 11, pp. 2855 – 2860, 2008.

[4] Shuo Wang, Ming Jiang, XavierMorin Duchesne, ElizabethA. Laugeson, DanielP. Kennedy, Ralph Adolphs, and Qi Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604 – 616, 2015.

[5] Ben Corden, Rebecca Chilvers, and David Skuse, "Avoidance of emotionally arousing stimuli predicts socialperceptual impairment in Asperger's syndrome," *Neuropsychologia*, vol. 46, no. 1, pp. 137 – 147, 2008.

[6] Kim M Dalton, Brendon M Nacewicz, Tom Johnstone, Hillary S Schaefer, Morton Ann Gernsbacher, Hill H Goldsmith, Andrew L Alexander, and Richard J Davidson, "Gaze fixation and the neural circuitry of face processing in autism," *Nature Neuroscience*, vol. 8, no. 4, pp. 519, 2005.

[7] Johanna Elisa Silberg, Ioannis Agtzidis, Mikhail Startsev, Teresa Fasshauer, Karen Silling, Andreas Sprenger, Michael Dorr, and Rebekka Lencer, "Free visual exploration of natural movies in schizophrenia," *European Archives of Psychiatry and Clinical Neuroscience*, Jan 2018.

[8] Smadar Levin, "Methodological Consensus in Smooth Pursuit Eye Movements: Workshop Contributions: Smooth Pursuit Impairment in Schizophrenia – What Does It Mean?," *Schizophrenia Bulletin*, vol. 9, no. 1, pp. 37–44, 01 1983.

[9] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. Le Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *ACM Multimedia Systems Conference (MMSys'19)*, June 2019.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 2106–2113.

[11] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[12] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 11, pp. 1254–1259, 1998.

[13] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, Oct 2018.

[14] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, March 2019.

[15] Timothy J Heaton and Megan Freeth, "Reduced visual exploration when viewing photographic scenes in individuals with autism spectrum disorder.," *Journal of Abnormal Psychology*, vol. 125, no. 3, pp. 399, 2016.

[16] Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan, "Scanpath modeling and classification with hidden Markov models," *Behavior Research Methods*, vol. 50, no. 1, pp. 362–379, Feb 2018.

[17] Sukru Eraslan, Victoria Yaneva, Yeliz Yesilada, and Simon Harper, "Web users with autism: eye tracking evidence for differences," *Behaviour & Information Technology*, vol. 0, no. 0, pp. 1–23, 2018.

[18] Ulf Rosenhall, Elisabeth Johansson, and Christopher Gillberg, "Oculomotor findings in autistic children," *The Journal of Laryngology & Otology*, vol. 102, no. 5, pp. 435439, 1988.

[19] Yukari Takarae, Nancy J. Minshew, Beatriz Luna, Christine M. Krisky, and John A. Sweeney, "Pursuit eye movement deficits in autism," *Brain*, vol. 127, no. 12, pp. 2584–2594, 10 2004.