Non-intrusive and calibration free visual exploration analysis in children with autism spectrum disorder

Dario Cazzato & Francesco Adamo  
*Faculty of Engineering, University of Salento, Lecce, Italy*  
Giuseppe Carlo Palestrra  
*Department of Computer Science, University of Bari, Italy*  
Giulia Criefaci, Paola Pennisi & Giovanni Piologgga  
*Institute of Clinical Physiology, National Research Council of Italy, Messina, Italy*  
Liliana Ruta  
*Department of Developmental Neuroscience, Stella Maris Scientific Institute, Calambrone, Pisa, Italy;*  
Marco Leo & Cosimo Distante  
*National Institute of Optics, National Research Council of Italy, Arnesano (LE), Italy*

**ABSTRACT:** Assistive technology is a generic system that is used to increase, help or improve the functional capabilities of people with disability. Recently, its employment has generated innovative solutions also in the field of Autism Spectrum Disorder (ASD), where it is extremely challenging to obtain feedback or to extract meaningful data. In this work, a study about the possibility to understand the visual exploration in children with ASD is presented. In order to obtain an automatic evaluation, an algorithm for free gaze estimation is employed. The proposed gaze estimation method can work without constrains nor using additional hardware, IR light sources or other intrusive methods. Furthermore, no initial calibration is required. These relaxations of the constraints makes the technique particularly suitable to be used in the critical context of autism, where the child is certainly not inclined to employ invasive devices. In particular, the technique is used in a scenario where a closet containing specific toys, that are neatly disposed from the therapist, is opened to the child. After a brief environment exploration, the child will freely choose the desired toy that will be subsequently used during therapy. The video acquisition have been accomplished by a Microsoft Kinect sensor hidden into the closet in order to obtain both RGB and depth images, that can be processed by the estimation algorithm, therefore computing gaze tracking by intersection with data coming from the well-known initial disposition of toys. The system has been tested with children with ASD, allowing to understand their choices and preferences, letting to optimize the toy disposition for cognitive-behavioural therapy.

1 INTRODUCTION

Autism is a neurodevelopmental disorder strongly characterized by deficits in social interaction and impaired understanding of the mental states of others (Siegel, Vukicevic, & Spitzer, 1990), a dysfunction that persists even in people with autism who have IQs in the normal range. Autistic Spectrum Disorders (ASD) is no longer a rare impairment with an extremely poor prognosis, but a multi-factorial disorder whose qualitative impairments can be mitigated with early intensive intervention. Worldwide population prevalence is about 1%. Early diagnosis of ASD is a central topic of research in developmental psychology. Signs of autism are not reliably present at birth, but emerge through a process of diminishing, delayed, or atypical development of social-communication behaviours, starting between the ages of 6 and 12 months. Intervention and support should be individualized and, if appropriate, multidimensional and multidisciplinary (Lai MC et al., The Lancet, 2014).

Aesthetic preferences of subjects with ASD are very different (Bogdashina, 2003). The preferred toy of a child can be intolerable for another child. Often, to understand whether an object is attracting or not the attention of an autistic child can be impossible because autistic children can use gestures or vocalization to express their needs by protoimperative gestures, but usually they don’t communicate objects of shared interest by protodeclarative gestures (Baron-Cohen, S., 1989; Curcio, F., 1978). Also in preverbal children, absence of protodeclarative gestures is considered a discriminating items between infants who had later been diagnosed with autism and typically developing infants (Clifford et al. 2007). Baron-Cohen et al. (1996) showed that absence of protodeclarative pointing, gaze-monitoring and pretend play in children 18 months old, in 83.3% of cases predicts a future diagnosis of autism. In Modified Checklist for Autism in Toddlers, protodeclarative gestures are considered a key behaviour (Robins et al. 2009).
Autistic subjects show atypically gaze behaviours: in real social situations, subjects with ASD show reduced salience of eyes and increased salience of mouths, bodies and objects (Klin et al. 2002; Shic et al. 2011) and look significantly less at the partner in dyadic interaction rather than typically development children (Noris et al. 2012). In artificial social situations, (for example when they see a cartoon or a movie with social actors, or in static photos, or in virtual reality, etc.) they attend to characters’ faces for less time that typically development subjects (Riby et Hancock 2009; Riby et Hancock 2008) and show less the centre of subjects’ faces rather than control group (Trepagnier et al. 2002). Visual atypical behaviours also affect the way in which subjects with ASD observe the scene; Shic et al. (2011) showed that children with ASD, in comparison to typically development subjects, are more attracted by background object. Many studies show an abnormal exploration of object stimuli (Ozonoff et al. 2008); the work of Sasson et al. (2008), for example, showed that, compared with typically development subjects, subject with ASD are more perseverative in fixations of details of images and that their fixations are more detail oriented, while Hutman et al. (2012) showed that infants later diagnosed with ASD are more attentive to non-social stimuli during interactions with an unfamiliar experimenter and that they shift attention among visual targets less frequently than high and low-risk typically developing children. By a simple preferential looking task, Pierce et al. (2011) showed that 40% of the autistic cross sample spent greater than 50% of viewing time fixating dynamic geometric images rather than dynamic social images. Several experimental protocols for the diagnosis and understanding of developmental disorders make use of video footage analysis to measure such elements as response time, attention changes and social interaction (Noris et al., 2007). Many approaches to technology-enhanced intervention rely on educational methods shown to result in good outcomes and can be used to specify design principles needed for engineering successful technology-enhanced intervention tools (Tentori M, Hayes G, 2010).

Considering the scarcity of such specialized medical practitioners compared to the volume of patients, along with considering the high-cost involved, there is a strong motivation for investigating technological alternatives to solve the problem (Sivalingam et al., 2012). From the other side, computer vision technology has a unique opportunity to impact that study of children’s behaviour, by providing a means to automatically capture behavioural data in a non-invasive manner and analyse behavioural interactions. Computational sensing and modelling techniques can play an important role in the capture, measurement, analysis, and understanding of human behaviour; this research area is called Behaviour Imaging.

In clinical settings, assessments of behaviour are typically based on the direct observation of a child by an experienced clinician. The ability to automatically measure behavioural variables using computer vision-based sensing could be valuable in enabling the collection of behavioural data on a large scale without requiring substantial human effort. Behaviour imaging technology can play several roles in support of a screening instrument. It can provide cost-effective tools for managing large collections of video and other data sources recorded during screening sessions. In particular, it can enable summarization, content-based retrieval, visualization, and comparison of observational data across populations and over time, to an extent that is not feasible using conventional manual methods (Rehg, 2011).

In this work, a study about a possibility to understand the visual exploration in children with ASD is presented. In order to obtain an automatic evaluation, an algorithm for free (i.e. without constrains nor using additional hardware, IR light sources or other intrusive methods) gaze estimation is employed. Furthermore, no initial calibration is required. In particular, the technique is used in a scenario where a closet containing specific toys, that are neatly disposed from the therapist, is opened to the child. After a brief environment exploration, the child will freely choose the desired toy that will be subsequently used during therapy. The video acquisition have been accomplished by a Microsoft Kinect sensor hidden into the closet in order to obtain a depth image that can be processed by the estimation algorithm, therefore computing gaze tracking by intersection with data coming from the well-known initial disposition of toys. The rest of the manuscript is organized as follows. Section 2 introduces the proposed method in terms of employed system and selected participants. Section 3 explores the experimental setup and the sensitive data acquisition phase, necessary to evaluate results off-line, and results are presented, while Section 4 discusses them. Finally, Section 5 concludes the paper.

2 METHODS AND PARTICIPANTS

The employed solution is based on the work of (Cazzato et al. 2012). A block diagram of the proposed method is showed in Fig. 1.

Fig. 1: A block diagram of the proposed method.
Method’s workflow is the following: first of all, images are acquired from a depth sensor, taking into account both RGB and depth channels. A search for a face over the input image is then carried out on the RGB image. A detected face is then tracked over the time, and a parameterized and people independent 3D face mask is subsequently overlapped on the human face. The face tracker is based on the Active Appearance Model (AAM), while the alignment is performed using Iterative Closest Point (ICP). The depth information allows to reconstruct the 3D positions of all of the tracked points and, therefore, to estimate the head pose in terms of yaw, pitch and roll angles. Starting from the average point between the two detected three-dimensional positions of the eyes and using the head pose information, a gaze track is finally estimated as well as its intersection with a vertical plane with regard to the ground and passing from the centre of the sensor, representing the observed point. Our gaze estimation method works as follows. First of all, the 2D position of the detected eye centre points are taken from the face mask. Note that small occlusions are handled, and the eye centre point is always estimated in the used model when the overlapping with the face successes. After that, the average value is taken, in order to translate the gaze vector to the right position. This way, the proposed method can be easily extended to compute the (possible) intersection with an object of interest in the plane. In this case, the virtual plane with lying objects is modelled with all the possible bounding boxes of the objects of interest. Therefore, the algorithm is able to determine if the gaze track ray has an intersection with a particular bounding box or not, determining which object in which frame have been observed.

About participants, the sample consisted of three boys (mean age: 50.3 months, range: 42-55 months), two with Autism Spectrum Disorder and one with Autistic Disorder, recruited at the Pervasive Healthcare Center of the Institute of Clinical Physiology of the National Research Council of Italy in Messina. The diagnostic assessment was based on the Autism Diagnostic Observation Schedule, Griffith’s Developmental Scale, Vineland Adaptive Scale, McArtur Language Test and Child Behaviour Checklist (CBCL).

3 DATA ACQUISITION AND EXPERIMENTS

Our method has been employed in a scenario where a closet containing specific toys, that are neatly disposed from the therapist, is opened to the child. Fig. 3 shows the disposition of the toys into the closet. The closet has size of 85x195 cm, but only three sectors (in the bottom) have been used, considering the age of the involved children. Thus, the useful part has a size of 85x117cm.

In Fig. 3, 9 Areas Of Interest (AOI) are designed. They have been manually computed, since a static disposition has been employed. Each AOI has a size of 28,3X39 cm. They are labelled from 1 to 9, and to each one of them correspond a specific toy. Exception, the cell number 2 will represent the Kinect area. Gaze data was extracted for each of these AOI. Fixations can be aggregated at many different levels, resulting in a wealth of psychophysical measures. In particular, the following parameters have been taken into account:

- Fixation count: the number of fixation in a specific AOI. For our system, a fixation of a toy is considered to be consistent if at least 15 consecutive frames present an hit on the square in exam;
- First fixation: the first consistent AOI looked by the patient;
- Sequence: the order of gaze hits into the AOIs;
- Selected Toy: the final choice of the patient;
- Most viewed toy: the cell with the maximum number of hits.
After a brief environment exploration, the child freely chooses the desired toy that will be subsequently used during therapy. Concerning the face of the child, video acquisition has been accomplished by a Microsoft Kinect sensor hidden into the closet in order to obtain a depth image that will be processed by the estimation algorithm, that will compute gaze tracking by intersection with data coming from the well-known initial disposition of toys. Depth image is necessary to estimate the head pose and to know the X,Y,Z world coordinate of the detected facial points. The green LED of the sensor has been darkened, to avoid to be appealing to the child and to invalidate our experiment. The depth sensor was connected on an Ultrabook Intel i3 CPU @ 1.8GHz with 4 GB of RAM, work at 30fps. The usage of our technique on a common Ultrabook was made in order to facilitate for raw installations like in the topmost shelves of the closet. This is a very encouraging result, since it allows to provide the useful data in real-time.

All recordings were suitable for off line analysis with our algorithm. For this experiment, one well-known disposition of the toys is used, and a patient gaze track is firstly computed and projected on the image plane; in a further step, its intersection with the toys’ AOIs have been computed. For each patient, these values have been recorded. Furthermore, these values have been elaborated, and results are summed up in Table 1.

Table 1. Gaze measures during tasks.

<table>
<thead>
<tr>
<th></th>
<th>Child 1</th>
<th>Child 2</th>
<th>Child 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixation Count</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Sequence</td>
<td>5-2-5-7</td>
<td>2-5-6</td>
<td>5-2-3-5-2-1</td>
</tr>
<tr>
<td>First Fixation Cell</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Selected Toy</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Most Viewed toy</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

2D visual exploration on a plane has been tracked over time, and this datum had been related with the chosen toy by means of the belonging bounding box. Fig. 4 shows the x and the y visual in an ordered sequence of frames where detection occurred, while Fig. 5 shows the computed hit map for the same patient during the interaction. The system was tested with three children, showing a completely different behaviour; in particular, the system tracked the first child only few frames, due to his very fast and out of field movements. The child selected a toy after a very short exploration, the first fixation was in the AOI 5 and the most viewed toy was in the AOI 2. There are no relations between what he watched most and the selected toy. The second child has a lower number of fixations than the other children. The selection went toward the last toy observed, in the opposite direction in respect to the toys seen previously. In this case, only with the usage of our system is possible to capture details and, in this case, to understand the real will of the child. Finally, in the third video, the system track different interval of the visual exploration, observing a large exploration of toys before concentrating on a point and taking the corresponding toy. This child has a higher number of fixations than the other children. The selected toy was the most viewed, He spent more time to see the chosen toy than the other. Figures 6 and 7 show the hit maps for the person #2 and #3, respectively.
DISCUSSION

In 2008, a case study reported by Vismara et Rogers showed a reduction in autistic symptoms thanks to a novel paradigm for treatment of autism based on an early intervention: the Early Start Denver Model (ESDM). Dawson et al. (2010) showed the efficacy of the ESDM for children with autism in improving cognitive and adaptive behaviours and reducing severity of ASD diagnosis. Rogers and Dawson (2010) emphasize the role of motivation to treat children with the ESDM. To motivate children they recommend to put subjects in a situation where there are many age-appropriate toys and then observe what children do. In this situation, children will probably watch toys more interesting for them, however the absence of protodeclarative gestures in children with autism (Baron-Cohen, S., 1989; Curcio, F., 1978) can be a problem for the therapist. In this contest, gaze estimator can be a good instrument to understand where children have focused its attention. Eye tracking is a useful methodology for investigating spontaneous priorities and patterns of attention. Relative priorities in attention can be inferred by analysing which regions of a stimulus a participant looks at, for how long the regions are looked at and when during the course of viewing the various regions are looked at.

This non-invasive method can facilitate our understanding of underlying cognitive processes involved in object perception and the more complex challenges related to social functioning in ASD (Boraston and Blakemore 2007). Knowing where a subject is looking provides a wealth of information regarding his motivations, expectations, and innate preferences. For this reason, eye tracking has become a standard tool for cognitive and psychological investigation. However, though the direct examination of the scan path obtained by eye-tracking systems tells us where individuals are looking, it does not tell us why they choose to look at different locations, and, of course, it does not tell us what these differences mean. In order to answer the more difficult question, that of how scan patterns should be interpreted, more sophisticated methods for distilling information from raw visual trajectories are required.

Before concluding, a discussion that considers the differences between a commercial eye tracker and the proposed system is reported. First of all, all the parameters of interest during a visual exploration of the patient have been taken into account. About them, it was possible to obtain a precise estimation of the number of fixations, the sequence of points of interest and where the attention goes at first. Two parameters that can also be estimated, but further development would be necessary, are the time of permanence and the average fixation length. Although it could be easily extracted adding a temporization on the detected frames, there were still a number of misdetection that could falsify the estimation, since we cannot automatically detect what happens in that frames (or it could be rebuilt by interpolation, but its value would not be reliable for the particular application context in exam). For this reason, we decided to do not report these values. From the other hand, advantages of the proposed systems are that the patient was able to participate to the session in a completely ecological way, i.e. without requiring any initial calibration, nor constraints about its position. Moreover, the proposed system is considerably inexpensive compared to commercial eye tracker solutions. This was a very encouraging results, that could provide a new inexpensive technique for evaluating focus-of-attention in autistic children.

Fig. 5 The computed hit map for person #1.

Fig. 6 The computed hit map for person #2.

Fig. 7 The computed hit map for person #3.
5 CONCLUSIONS

Computer vision technology has a unique opportunity to impact that study of children’s behaviour, by providing a means to automatically capture behavioural data in a non-invasive manner and analyse behavioural interactions between children and toys. Eye tracking behavioural assessments may provide spatial information than face-to-face assessments. These tools aid the clinician in the behavioural assessment task by providing accurate and objective measurements.

The early detection of developmental disorders is key to child outcome, allowing interventions to be initiated that promote development and improve prognosis. Research on ASD suggests behavioural markers can be observed late in the first year of life. Many of studies involved extensive frame-by-frame video observation and analysis of a child’s natural behaviour. Although non-intrusive, these methods are extremely time-intensive and require a high level of observer training. With the goal of aiding and augmenting the visual analysis capabilities in evaluation and developmental monitoring of ASD, we proposed computer vision tools to observe specific behaviours related to ASD elicited during toy selection tasks, providing both new challenges and opportunities in video analysis.

6 ACKNOWLEDGEMENTS

This work has been supported by the "2007-2013 NOP for Research and Competitiveness for the Convergence Regions (Calabria, Campania, Puglia and Sicilia)" with code PON04a3_00201.

7 BIBLIOGRAPHY


