Towards Automated Video Analysis of Sensorimotor Assessment Data

Ana B. Graciano Fouquier[1], Séverine Dubuisson[2], Isabelle Bloch[3] and Anja Klöeckner[4]
[1]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France*
[2]*Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7222, ISIR, F-75005, Paris, France*
[3]*Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France*
[4]*Dept. of Child and Adolescent Psychiatry, APHP, Hôpital Pitié-Salpêtrière, Sorbonne Universités, UPMC Univ Paris 06, Paris, France*
ana.graciano@lip6.fr, severine.dubuisson@isir.upmc.fr, isabelle.bloch@telecom-paristech.fr

Keywords: Video Analysis, Behavioral Imaging, Autism Spectrum Disorders, Sensorimotor Assessment.

Abstract: Sensorimotor assessment aims at evaluating sensorial and motor capabilities of children who are likely to present a pervasive developmental disorder, such as autism. It relies on playful activities which are proposed by a psychomotrician expert to the child, with the intent of observing how the latter responds to various physical and cognitive stimuli. Each session is recorded so that the psychomotrician can use the video as a support for reviewing in-session impressions and drawing final conclusions. These recordings carry a wealth of information that could be exploited for research purposes and contribute to a better understanding of autism spectrum disorders. However, the systematic inspection of these data by clinical professionals would be time-consuming and impracticable. In order to make these analyses feasible, we discuss a computer vision approach to prospect behavior information from the available visual data acquired throughout assessment sessions.

1 INTRODUCTION

The study of autism spectrum disorders (ASD) assisted by computer vision tools has recently emerged in the literature (Rehg, 2011; Porayska-Pomsta et al., 2012) and revealed a new potential research field. This is partly motivated by increasing governmental efforts in different countries to support studies on ASD, since nowadays these disorders are more frequently diagnosed and their lifetime treatment implies a substantial public budget. Therefore, better understanding of these disorders should impact the way we deal with them, and it could lead to earlier diagnosis, as well as to more effective and less costly treatments.

Visual data are often a useful source of knowledge for ASD diagnosis and research. For instance, home movies have been traditionally used as a means of evaluating ASD hypotheses (Zakian et al., 2000; Bernard et al., 2006; Wendland et al., 2010). The uncontrolled environments depicted in these movies could be used to detect or verify numerous aspects of ASD if they were explored automatically. Also, data obtained from various ASD diagnostic protocols (e.g. AOSI, MCHAT, sensorimotor assessment) could be exploited to test hypotheses or reveal hidden patterns over groups of children. However, the amount of available data makes it virtually impossible for a human expert to go through it all in pursuit of some target behavior or signs. Therefore, the design of computer vision tools targeted at these studies should provide a means of dealing with the large amount of recorded data in a more efficient way. Ultimately, they may contribute to a better understanding of autism spectrum disorders. In this position paper, we will discuss these potentialities through a case study.

Related Work. Recent efforts have been made towards these automated analyses. In (Hashemi et al., 2012), the authors present a semi-automatic computer-vision-based methodology to analyze three activities from the AOSI (Autism Observation Scale for Infants) diagnostic protocol and to study postural patterns related to autism. These activities evaluated the child's ability to share attention, to perform a visual pursuit of an object, and to switch attention from one stimulus to another. To perform these experiments, the authors implemented a set-up to acquire video data from patients aged 6-15 months. These videos were not part of the usual protocol and thus had to be obtained for this study purposes. A fixed camera was used to record the target activities, thus the perspective of the scene remained the same dur-

ing each activity and for each patient. The vision-based analysis relied on geometrical measurements estimated through the triangle formed by the patient's nose, the left ear and eye, which were always visible in the videos. The reported automated results corroborated with what clinician's usually conclude, which proved that specialized automated methods can be used as auxiliary tools for processing this type of information.

The work from (Rehg et al., 2013) proposes a multi-modal (audio and video) method for analyzing fine-grained actions and behavior manifested during a session of the Rapid-ABC diagnostic protocol. For this study, the authors proposed a complete setup to acquire short videos of the interaction between a child and the clinician during a single activity of the Rapid ABC. The setting was composed of several fixed cameras (2 frontal views, 8 side views, 3 overhead views), an overhead view Kinect©, as well as microphones and electrodermal sensors. This system produced an annotated public database (*MMDB*) of 160 videos with children aged 15-30 months. These data were parsed using speech recognition technology in order to segment the activity into its main stages. Then, the visual data were put together to evaluate the degree of engagement between the child and the clinician. Techniques for smile and gaze detection, object recognition and tracking, event detection and feature extraction were adopted in order to predict this engagement. The study also revealed various problems in the computer-assisted study of dyadic relations. Again, the encouraging results point to the feasibility of automated analysis of ASD data, and bring hope for a better understanding of these disorders.

Position Paper Objectives. In this position paper, we shall discuss how automated video analysis can help the study of videos obtained during sessions of sensorimotor assessment. This diagnostic method was proposed by A. Bullinger (Bullinger, 2006) to evaluate a child who possibly presents a pervasive developmental disorder (PDD), such as autism, from an evolutionary point of view regarding sensory and motor capabilities. A typical sensorimotor assessment session is always recorded for later use as a review tool for verifying the psychomotrician's in-session impressions. Unlike the aforementioned works, these data were part of the protocol and were recorded in an uncontrolled environment by an assistant psychomotrician using a single personal camcorder manipulated freely. This fact implies that our system must cope with technical issues that are still unresolved, such as dealing with important zooming and occlusions without any 3D information, as well as with motion, low image quality and artifacts. Also, the children age range varies from toddlers to teenagers, which also supposes a flexible system in order to cope with their inherent variability when modeling the possible course of actions.

From a computer vision perspective, automatic analysis of these recordings based on visual data entails object detection, segmentation, classification and tracking. Although there is a vast literature on each of these problems, the difficulties due to the data acquisition under an uncontrolled setting lead to failure of various existing methods. Also, the augmented dimension of interpreting cognitive and social signals is a challenge, since it requires the computational modeling of numerous types of human behavior and social concepts and the development of specific social signal measures. It is important to mention that the proposed automated analysis does not aim at replacing the clinician's diagnosis, but rather to provide ways to inspect the data *a posteriori* and to verify the clinician's research hypotheses and impressions about seemingly common patterns of behavior among PDD patients.

In the remainder of the paper, all these points shall be discussed as follows. Section 2 describes a typical assessment session and details one of its activities, which shall be the focus of our automated analysis. Section 3 presents the available video data for this activity and assesses their challenges from a computer vision and video analysis perspective. The clinical hypotheses verifiable from these excerpts and their computational representation are discussed in Section 4. Preliminary results on object segmentation and tracking as a first step for the automated video analysis are given in Section 5. Finally, Section 6 brings conclusions and future perspectives on the problem.

2 OVERVIEW OF THE SENSORIMOTOR ASSESSMENT SESSION

As stated in Section 1, sensorimotor assessment is a diagnostic protocol for evaluating sensorimotor integration in patients likely to suffer from a PDD. Each session consists of a series of activities that observe the responsiveness of the child to distinct stimuli (visual, tactile, motor, auditory) presented by a psychomotrician professional (Jutard et al., 2009; Kloeckner et al., 2009). It also assesses the child's ability to share attention, to engage in social activities, to understand and to perform a given task.

The whole session lasts about one hour and a half and is entirely recorded. Since the psychomotrician

must be entirely committed to the assessment, he/she cannot take notes during the session. Thus, the resulting video serves as a record of observable assessment data, which the clinician may later use to review his/her in-session impressions and to write a final assessment report. The video also serves as a means of explaining the psychomotrician's observations to the child's parents. An auxiliary psychomotrician always takes part in these sessions. This clinician is responsible for recording the activities with a single camcorder which can be freely operated. The objective is to capture the global progress of the test, as well as to register in detail individual gestures or expressions that might be remarked and considered relevant.

As a first step towards our automated analysis system, we will focus on a single activity of the assessment. We shall refer to it as the "grab the stick test". For this particular test, the child is invited to sit on a chair and the psychomotrician usually kneels down in front of him/her. Then, the psychomotrician sequentially presents to the child a series of identical wooden sticks from different starting points in space and in distinct orientations. Ideally, the child is expected to grab one stick at a time with one hand and keep the previous ones in the other hand. When the activity is completed, the child is asked to store the set of sticks in a pencil case and close its slide fastener. Figure 1 shows a few sample frames extracted from the video passage containing this activity.

Figure 1: Sample frames from the "grab the stick test".

The choice of this activity was not casual. This test offers clinical evidence for a set of sensorimotor handicaps. Apart from evaluating cognitive skills through task understanding and learning (child's response to test instructions), the clinical purposes of this activity assess oculomanual coordination (spatial perception and localization of stimuli, followed by hand movement), median line crossing (capability of communication between the right and left halves of the body), tactile dexterity and rhythm (movement frequency, elapsed time between grabbed sticks), and grip quality (intensity, adequacy).

3 VIDEO DATA ANALYSIS

The examples in Figure 1 depict the usual environment where the sensorimotor assessment sessions are performed. The setting consists of a bright room where all the tools for the sensorimotor assessment are kept. For this activity, the auxiliary psychomotrician usually records the activity from a side perspective in relation to the other professional and the child. A chair is always required, and sometimes cushions, feet-support, or a Vichy-textured board are deployed to make the setting more comfortable for the child. In all situations, the child is sitting in front of the psychomotrician and the assistant must avoid interfering at all costs.

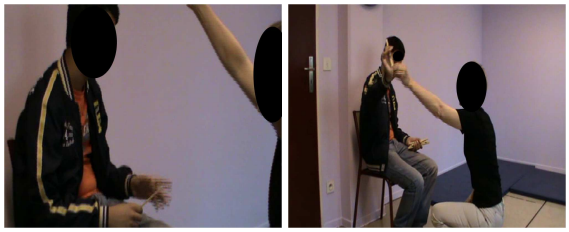
We obtained a total of nine video recordings from sensorimotor assessments. Three of the videos were from female patients, while the remaining ones were from male patients. Each full video was manually cropped so that the excerpts analyzed contained uniquely the "grab the stick test". This reduces the total video time from approximately 1,5 hours to a 1- to 5-minute excerpt, which is more reasonable for performing object tracking. Table 1 presents a summary of these videos by listing the patient's gender and age, the visible people (actors) present in the excerpt besides the patient and the psychomotrician, the suspected diagnosis, and the technical difficulties inherent to the video data or to the analysis of the test.

As it can be seen, total or partial occlusions are one of the most recurrent issues among all videos (Figure 2 (b)). These are often due to the bad adjustment of the camera visual field, which is sometimes insufficient to capture all the spatial locations in which the sticks appear and where the child grabs them. Furthermore, since the shooting is usually lateral, it may capture the psychomotrician from an angle which partially occludes the child or the activity. Another common issue results from fast zooming (Figure 2 (a)). Due to clinical needs, it is sometimes applied to focus on a localized expression or reaction of the child, which abruptly changes the scene and leaves most important elements out of sight. Zooming is also a source of temporary stick occlusion and disappearance.

Besides these, sticks and hands may suffer from motion blur at times due to insufficient frame rate, which makes it harder to segment and recognize them.

Table 1: Summary of Sensorimotor Video Contents. For each video at our disposal, the table presents relevant information such as the patient's gender (Female/Male), age (in years), people appearing in the video besides the patient and the psychomotrician, the likely diagnosis, and the events that occur during the task that can make its automated analysis difficult.

Gender	Age	Actors	Suspected Diagnosis	Video challenges
F	3	Mother	Anorexia and slight ASD	Sticks out of view; sticks occluded by hands/other objects; task repeated twice
F	7	Father	Schizophrenia	Sticks/hands out of view; child plays with sticks
F	12	Parents	ASD	Interaction between the child and each parent; zooming; complex behavior due to global impairment (wheelchair)
M	2.5	Parents	ASD	Child sitting on mother's knees; sticks out of view; sticks occluded by hands/other objects;
M	3.3	N/A	ASD	Child plays with sticks for a while and places them on the floor; presence of Vichy-textured board; sticks out of view; sticks occluded by hands/other objects;
M	5	Father	ASD	Sticks out of view; sticks occluded by hands/other objects; zooming
M	10.5	N/A	ASD	Presence of child's puppets; sticks occluded by hands/other objects; zooming, camera instability
M	14	Father	ASD	Sticks occluded by hands/other objects; zooming
M	16.5	N/A	ASD	Sticks occluded by hands/other objects; zooming



(a) (b)

Figure 2: Typical technical problems encountered in videos of the "grab the stick test": (a) blurred elements (the child's hand), out-of-sight objects (the stick, the psychomotrician's hand) and zooming, (b) partial occlusion due to side perspective.

Undesirable motion is also caused by the fact that the video is captured from a camera manipulated freely. This makes it harder to successfully adopt background subtraction techniques to isolate foreground elements. Also, when parents are present, the number of human interactions increases the number of occlusions and makes object segmentation and tracking more difficult. Finally, the variability in patients responses adds to the challenge of modeling behavior and courses of actions for this activity.

4 AUTOMATED INSPECTION AND EVALUATION OF AUTISTIC SIGNS

As described in Section 2, the "grab the stick test" is particularly useful for screening the child's learning and developmental limitations regarding various aspects. For the automated analysis though, we shall focus on a complementary evaluation related to task understanding, as well as to signs of sensorimotor difficulties during task completion.

Intuitively, the psychomotrician can verify if a child understands the activity by waiting for his/her reaction to the first stick shown. Therefore, the time a child takes to respond can be related to his/her understanding of the task. In addition to this research hypothesis, the clinician could verify whether a child reacts faster and automatically after grabbing the first sticks, which could reveal an underlying cognitive process. The time-of-response variability throughout the activity may be an index of this learning process. Yet another research hypothesis concerns the signs of sensorimotor development impairment. From the point of view of sensorimotor analysis, if the child produces atypical mouth gestures while trying to reach out for the stick or while bringing it closer to his/her body, then there may be a motor trouble that the child tries to compensate with the mouth movement as if it were a support.

These research hypotheses may be verified by either measuring the respective elapsed times, or observing simultaneous stick grabbing and mouth movement, based on the available "grab the stick test" video recordings for all patients. Their automated analysis through computer vision implies the following low-level problems: detection of sticks, hands and mouth; recognition of the patient's left and right hands and the psychomotrician's hands; tracking of sticks and hands. Subsequently, these elements will compose the building blocks for modeling a typical course of action of the "grab the stick test". We will

suppose that the analysis will only be valid whenever all these elements are within the limits of the depicted scene in each video frame.

To represent the course of the activity, we specify the following key action patterns to look after in the video:

- a new stick appears in the scene: to facilitate this task, the first stick is manually detected/segmented; the subsequent ones must be detected automatically by recognizing an intersection or adjacency between the region recognized as the psychomotrician's hand and the region detected as a stick;
- the child grabs a stick: this pattern is detected by searching for an intersection or adjacency among the regions detected for the child's and psychomotrician's hands and the stick;
- the child moves his/her hand: if a region recognized as one of the child's hand presents considerable displacement from one frame to another, this should trigger the tracking of this hand;
- unusual mouth gesture: the frame region detected as the mouth must be recognized as an atypical gesture; a computational model of these unusual gestures must be created so that the recognition process can disambiguate them from acceptable patterns such as smiling, speech, and yawning.

The first three patterns serve as markers for the succeeding step, which aims at verifying the hypotheses of learning and measured reaction times. The last three patterns can be put together in order to verify the hypothesis of sensorimotor impairment.

5 PRELIMINARY RESULTS

As discussed in Section 4, the first stage of our automated video analysis consists of low-level detection, recognition and tracking of target elements. Our preliminary experiments aim at evaluating the research hypotheses related to task understanding and learning. To this aim, we explore available *a priori* knowledge concerning the videos at our disposal in order to detect and track hands and sticks.

Color Cues. Hands naturally relate to skin color. Skin detection in images has already been discussed by a series of works (Albiol et al., 2001; Jones and Rehg, 2002) and skin colormaps are widely available. Thus, we use this cue in the hand detection step. The original sticks are all of the same color. In spite of variations due to illumination effects in the videos, an empirical analysis of their hue-saturation-value (HSV) histograms has shown that they often fall into a rather well-defined color range, which supports the use of color features for stick detection. Curiously, the color of the sticks also falls into the skin range, which allows us to use the same color filter for both hands and sticks.

Shape Cues. All sticks present the same shape and size. Although perspective transformations might change their appearance throughout the videos, their linear geometry is valuable information to be exploited. We seize this information by adopting the probabilistic version of the Hough Transform (Matas et al., 2000) to detect lines. We apply this filter over the image produced by a morphological opening of the skin detector result, then we filter the results by line size in order to remove spurious lines.

Stick Tracking. We track the sticks using a particle filter model inspired by the one presented in (Pérez et al., 2002). The state space consists of the position (2D image coordinates) and the scale of the sticks. We adopt a random-walk model to represent the system dynamics. Finally, the likelihood function is computed with respect to a reference stick color model based on the HSV color space. In order to initialize the position state variable, we manually define a mask around the stick upon the first frame where it appears. We have successfully tracked the first stick up to the moment the child grabs it and starts to retract his/her arm, as shown in Figure 3.

Figure 3: Sample results for the tracking of sticks through particle filtering.

Although these features have been used in our present experiments, others might as well be adopted.

6 CONCLUSIONS

Automated analysis of ASD clinical data shall help to reveal new findings concerning autistic disorders.

In this position paper, we discussed how computer vision and video analysis can be employed to evaluate video data from sensorimotor assessment, a diagnostic protocol for screening pervasive developmental disorders such as autism.

Our next steps aim at concluding the analysis of the "grab the stick test" as discussed hereby. This will allow us to measure the times of reaction for all sticks and estimate their variability for each patient. Then a global study of these results should help sustain or reject both task understanding and learning hypotheses. We shall also verify the relation between sensorimotor impairment and mouth gestures throughout the test. A set of common mouth gestures must be defined in order to distinguish them from unusual ones. One solution to this problem is to model mouth gestures through the Facial Action Coding System, an approach adopted in successful recent works (Mahoor et al., 2009; Senechal et al., 2013).

Future research shall also deal with video editing and other behavior analysis from video. The first category concerns the automatic detection of zooming and video passages that provide insufficient content for an analysis by either a human expert or computer vision tools, as it is the case when sticks or hands are not visible in a scene. The automatic segmentation of the assessment videos into passages corresponding to distinct sensorimotor activities shall also be exploited.

The second category covers the study of signs of fatigue and loss of attention during the assessment. This research shall evaluate the videos thoroughly in order to look for remarkable signs, such as yawning or torso relaxation and bending. The clinical motivation is to perceive common signs related to these manifestations, as well as to understand which stimuli might help the child to regain attention.

ACKNOWLEDGEMENTS

The videos used in this study were acquired with the informed consent of the parents, who agreed to the use of the data for educational and research purposes.

Ana B. G. Fouquier was the recipient of a Post Doctoral Fellowship provided by the Brazilian National Council for Scientific and Technological Development (CNPq).

REFERENCES

Albiol, A., Torres, L., and Delp, E. (2001). Optimum color spaces for skin detection. In *Proc. IEEE Int. Conf. Image Process.*, volume 1, pages 122–124.

Bernard, J. et al. (2006). Evolution de la fréquence des gestes chez des garçons avec autisme âgés de 1 à 3 ans par analyse de vidéos familiales. *Devenir*, 18(3):245–261.

Bullinger, A. (2006). Approche sensorimotrice des troubles envahissants du développement. *Contraste*, 2(25):125–139.

Hashemi, J. et al. (2012). A computer vision approach for the assessment of autism-related behavioral markers. In *Proc. IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7.

Jones, M. J. and Rehg, J. M. (2002). Statistical color models with application to skin detection. *Int. J. Comput. Vision*, 46(1):81–96.

Jutard, C., Kloeckner, A., Périsse, D., and Cohen, D. (2009). Intérêt de l'abord sensorimoteur dans les pathologies autistiques sévères II : illustration clinique. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, 57(2):160 – 165.

Kloeckner, A. et al. (2009). Intérêt de l'abord sensorimoteur dans les pathologies autistiques sévères I : introduction aux travaux d'André Bullinger. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, 57(2):154 – 159.

Mahoor, M. et al. (2009). A framework for automated measurement of the intensity of non-posed facial action units. In *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition Workshops*, pages 74–80.

Matas, J., Galambos, C., and Kittler, J. (2000). Robust detection of lines using the progressive probabilistic hough transform. *Comput. Vision and Image Understanding*, 78(1):119–137.

Pérez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *Proc. European Conf. Comput. Vision*, pages 661–675.

Porayska-Pomsta, K. et al. (2012). Developing technology for autism: an interdisciplinary approach. *Personal Ubiquitous Comput.*, 16(2):117–127.

Rehg, J. M. (2011). Behavior imaging: Using computer vision to study autism. In *IAPR Conf. Mach. Vision Appl.*, pages 14–19.

Rehg, J. M. et al. (2013). Decoding children's social behavior. In *Proc. IEEE Comp. Soc. Conf. Comput. Vision and Pattern Recognition*, pages 3414–3421.

Senechal, T. et al. (2013). Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience. In *Int. Conf. Automatic Face and Gesture Recognition*, pages 1–8.

Wendland, J. et al. (2010). Retrait relationnel et signes précoces d'autisme : étude préliminaire à partir de films familiaux. *Devenir*, 22(1):51–72.

Zakian, A. et al. (2000). Early signs of autism: A new study of family home movies. *L'Encéphale*, 26:38–44.