An Interactive Multimedia System for Treating Autism Spectrum Disorder

Massimo Magrini[1,2(&)], Ovidio Salvetti[1,2], Andrea Carboni[1,2], and Olivia Curzio[1,2]

[1] ISTI-CNR, Pisa, Italy
{Massimo.Magrini,Ovidio.Salvetti,
Andrea.Carboni}@isti.cnr.it,
[2] IFC-CNR, Pisa, Italy
Olivia.Curzio@ifc.cnr.it

Abstract. A system for real-time gesture tracking is presented, used in active well-being self-assessment activities and in particular applied to medical coaching and music-therapy. The system is composed of a gestural interface and a computer running own (custom) developed software. During the test sessions a person freely moves his body inside a specifically designed room. The algo-rithms detect and extrapolate features from the human figure, such us spatial position, arms and legs angles, etc. An operator can link these features to sounds synthesized in real time, following a predefined schema. The augmented interaction with the environment helps to improve the contact with reality in subjects having autism spectrum disorders (ASD). The system has been tested on a set of young subjects and a team of psychologists has analyzed the results of this experimentation. Moreover, we started to work on graphical feedback in order to realize a multichannel system.

1 Introduction

In recent years specific activity has been carried out for developing sensor-based interactive systems capable to help the treatment of learning difficulties and disabilities in children [1–3]. These systems generally consist of sensors connected to a computer, programmed with special software that reacts to the sensor data with multimedia stimuli. The general philosophy of these systems is based on the idea that even pro-foundly physically or learning impaired individuals can become expressive and com-municative using music and sound [4]. The sense of control which these systems provide can be a powerful motivator for subjects with limited interaction with reality.

While a great part of systems, like SoundBeam (www.soundbeam.co.uk), totally rely on ultrasonic sensors, our approach is mostly based on real-time video processing techniques; moreover, our solution makes also it possible to easily use additional sets of sensors (e.g., infrared or ultrasonic) in the same scene. The use of video-processing techniques adds more parameters suitable to localize exactly and detail all the human gestures we want detect and recognize. By using a custom software interface, the operator can link the extracted video features to sounds synthesized in real time, following a predefined schema.

The developed system has been experimented in a test campaign on a set of young patients affected by Low-Functioning Autism (autism spectrum disorder, ASD), in order to provide a personal increased interaction over the operational environment and to reduce pathological isolation [5]. Results were very positive and encouraging, as con-firmed by both clinical psychologist and parents of the kids. In particular, the therapists reported a positive outcome from the assisted coaching therapies. Indeed, this positive evolution was crucial to improve the motivation and curiosity for a full communication interaction in the external environment, thus affecting subjects' well-being.

In order to maintain the obtained benefits, we developed a simplified version of the system, based on Kinect, to be used at home by the parents and children. The present project program trains parents of children with autistic spectrum disorders using the DIR/Floortime model of Stanley Greenspan MD. Parents were encouraged to deliver 15 h per week of 1:1 interaction. Pilot studies [6, 7] suggested that this kind of models have potential to be a cost-effective intervention for young children with autism.

2 Autism

ASD is a neurodevelopmental disorder characterized by impaired social interaction and communication. It is a pervasive developmental disorder, characterized by a triad of impairments: social communication problems, difficulties with reciprocal social inter-actions, and unusual patterns of repetitive behavior [8]. Leo Kanner, a child psychiatrist [8], described it for the first time in 1943. An exhaustive description of this disorder in medical terms is beyond the scope of this paper.

Unfortunately, no medications can cure autism or treat its core symptoms, but rather can help some people affected feel better. A large part of the interventions focuses on behavioral approaches, of which the best known is the ABA (Applied Behavior Analysis) method [10], based on repetitive patterns and reinforcements. Other approaches follow instead the Developmental Individual Difference Relationship (DIR) model [11]. DIR acts at various levels of involvement, attempting a containing action against the central symptoms of autism according to the following guidelines:

(1) involvement against isolation (2) communication and flexibility versus rigidity and persistence (3) gestures against stereotypies and aggressive behaviors. In the design of our system we were inspired, even not strictly, by the DIR model.

3 The System

The system was developed in two different steps, using different technologies. The first version is based over video capture and processing, where all the body recognition routines are software-based. This system needed a controlled environment and the presence of a technician during each session. The second version relies on the use of Kinect v2, which solved some of the problems of the previous version. This version allows the tracking of full body movements in 3D space, has the peculiarity of being designed to be installed in the user's home, and provides an intuitive interface, to be easily used by the children's families.

3.1 Camera Based Version

The first version of the system has been based on real-time video processing. The software has been developed in C++ on an Apple Macintosh computer running the latest version of Mac OS X. A video camera is connected through a Firewire digitizer, the Imaging Source DFG1394, a very fast digitizer that allows a latency of only one frame in the video processing path. As output audio card we use the Macintosh internal one, sufficient for our purposes. A couple of TASCAM amplified loudspeakers com-pletes the basic system.

We used the Mac OS platform for its reliability in real time multimedia applications, thanks to its very robust frameworks: Core Audio and Core Image libraries permit very fast elaboration without glitches and underruns.

Finally, the system is installed in a special empty room, with most of the surfaces (walls, floor) covered by wood. The goal is to build a warm space which, in some way, recalls the prenatal ambient. All system parts, such as cables, plugs, and so on are carefully hidden as they could be potential elements of distraction. The ambient light is gentle and indirect, thus avoiding shadows that could also affect the precision of motion detection. Nevertheless, large changes in the environment light may affect the system's setup, so that the software needs some recalibration.

System Architecture. The implemented system (Fig. 1) is organized in several specialized coordinated modules. The core of the software is composed by the Sequence grabber, which manages the stream of video frames coming from the video digitizer, the Video processor, which performs realtime image elaborations, the Skeleton reconstructor, which analyses the frames and extrapolates gesture parameters, the Data mapper, responsible for transforming the detected gesture parameters into sounds parameters and finally the Sound synthesizer.

The biggest problem regarding the gesture control of sounds is latency, which is the delay between the gesture and the correspondent effect on the generated sound.

Our approach guarantees the minimum latency for the adopted frame rate, which is 40 ms at 25 FPS or 33.3 ms at 30 FPS.

Processing algorithm. In the first step of the algorithm each grayscale frame grabbed in real time from the video camera is smoothed with a Gaussian filter (fast computed using the CoreImage library). The output is then processed in one of two alternative operating modes: area-based or edge-based. In the area-based modality, the segmen-tation is performed considering the entire envelope (area) of the figure of the subject examined. In the edge-based modality, instead, an edge detection filter is applied to the image. The next step consists in a background subtraction technique computed to isolate the human figure from the ambient. In order to fulfill this task, each time the background changes, it has to be stored, area or edge based, with no human subject in front of the camera. If this background exists it is used in the following iterations of the algorithm and compared with the incoming frames containing the human figure using a dynamic threshold, obtaining a binary matrix. The average threshold used in this operation can be tuned by the operator in real time. It is not necessary to set again this sensitivity if the ambient light does not change. Finally, we apply an algorithm for removing unconnected small areas from the matrix, usually generated by image noise.

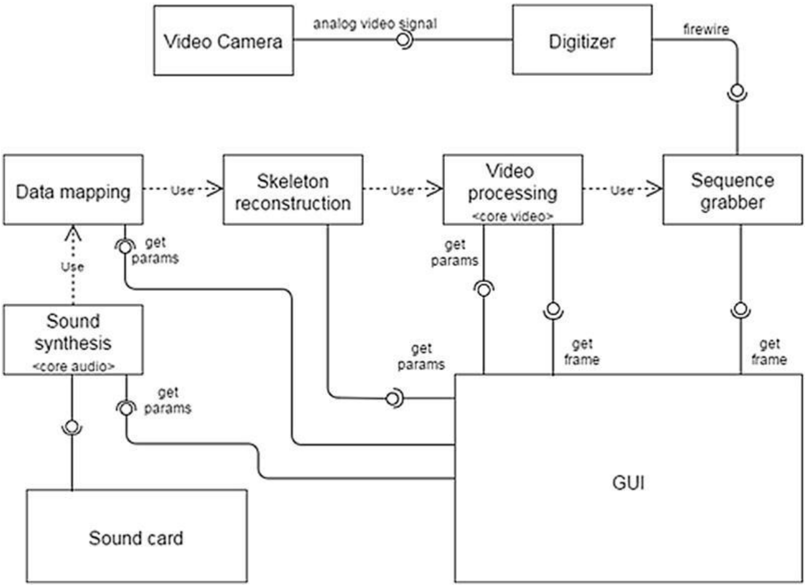
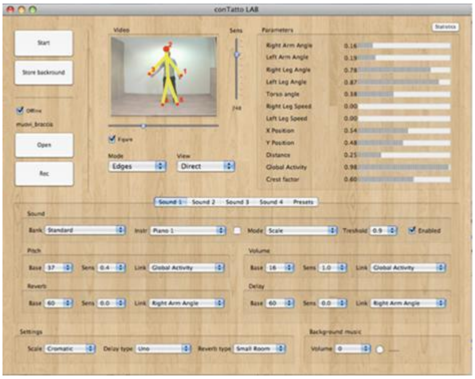


Fig. 1. System architecture.

The final binary image is then ready to be processed by the gesture tracking algorithm. The frame resolution is 320 240 pixels, full frame rate (e.g. 25 FPS) can be achieved because all the image filters are executed by the GPU. Starting from the binary raster matrix we apply an algorithm to detect a set of gesture parameters. This heuristic algorithm supposes that the segmented image obtained by the image elaboration pro-cess is a human figure, extracting data from it. This process starts searching a simplified model of the human figure, shown in the GUI (Fig. 3). Additional models for detailed parts of the body (face, hands) are currently under development, so that they can be used for more "zoomed" versions of the system. Starting from time dependent position of detected body joints we decided to compute the following parameters: Arm angles and speed (left/right), Leg angles and speed (left/right), Torso angle, Barycenter coordinates, Distance from camera.

The distance from the camera actually is just an index related to the real distance: it is simply computed as a ratio between the frame height and the detected figure maximum height. The leg speed is computed analyzing the last couple of received frames; it is useful for triggering sounds with "kick-like" movements. We also compute these two additional parameters: global activity, crest factor (Fig. 2).

The first is an indicator of overall quantity of movement (0.0 if the subject is standing still with no moments), while the second one is an indication of the concavity of the posture: (0.0 means that the subject is in standing position with arms kept along the body). Few optimizations are performed starting the frame analysis from an area centered in the last detected barycenter. Generally speaking, we tried to implement the detection algorithms in a very optimized way in order to maintain the target frame rate (25 FPS), minimizing the latency between gestures and sounds.



☑ Offline
muovi_braccia

☑ Figure

Mode: Edges
View: Direct

Right Arm Angle	0.16
Left Arm Angle	0.19
Right Leg Angle	0.78
Left Leg Angle	0.87
Torso angle	0.38
Right Leg Speed	0.00
Left Leg Speed	0.00
X Position	0.54
Y Position	0.48
Distance	0.25
Global Activity	0.98
Crest factor	0.60

Sound: Bank Standard, Instr Piano 1, □ Mode Scale, Treshold 0.9, ☑ Enabled

Pitch: Base 37, Sens 0.4, Link Global Activity

Volume: Base 16, Sens 1.0, Link Global Activity

Reverb: Base 60, Sens 0.0, Link Right Arm Angle

Delay: Base 60, Sens 0.0, Link Right Arm Angle

Settings: Scale Cromatic, Delay type Uno, Reverb type Small Room

Background music: Volume 0 ○

Fig. 2. Graphical User Interface

Sound Generation. The sound generation is based on the Mac OS CoreAudio library. The mapper module translates the detected features into MIDI commands for the musical synthesizers. There are four independent synthesizers: for each of them the sound’s parameters (pitch, volume, etc.) can be easily linked to the detected gesture parameters using the GUI. For example, we can link the Global Activity to the pitch: the faster you move the higher pitched notes you play. The synthesized MIDI notes are chosen from a user selectable scale: there’s a large variety of them, ranging from the simplest ones (e.g. major and minor) to the more exotic ones. As an alternative, it is possible to select continuous pitch, instead of discrete notes: in this way the linked detected features controls the pitch in a “glissando” way. Sound can be triggered in a “Drum mode” way, too: the MIDI note C played when the linked parameters reach a selected threshold. All these links settings can be stored in presets, easily selectable from the operators.

3.2 Kinect Based Version

The experimentation has proven the positive effects so, in order to maintain its benefits, we developed a more user-friendly version to be used at home, avoiding the need of specialized personnel.

The user interface of this home version is greatly simplified compared to the video camera one and simply presents a series of easily selectable, not editable, presets. The body detection algorithm benefits from the use of Microsoft Kinect v2 SDK, which not requires the critical camera settings of the first version.

Microsoft Kinect. The Microsoft Kinect is a line of well-known motion sensing input devices originally created for Xbox 360 and Xbox One video game consoles and then Windows PCs. This device enables users to control and interact with their console/computer without the need for a game controller, through a natural user inter-face using gestures. The first Kinect version was using a structured infrared light approach while V2 (the version we used) is based on the Time-of-Flight (ToF) principle. Using these technologies, the device can compute a depth map of the environment. The Microsoft

Kinect SDK libraries can process this depth map and extract the tridimensional coordinates of (up to) 26 joints of the human skeleton. Up to 6 skeletons can be tracked in real time. While Kinect V1 (which was considered during the development of the camera version) still suffered from large latency, the V2 partially solved this issue so that we finally decided to use it in the home version.

Software. This architecture of the SW, compared to the one described in Sect. 3.1, has the three modules Sequence Grabber Video Processor and Skeleton Reconstruction, replaced by routines developed with the Kinect SDK. Since this SDK provides only the spatial coordinates of the joints, we included a module for performing geometrical transformations on them, computing a set of features, basically the angles between body parts, that are invariant in relation to the body position and rotation. These features are then aggregated and mapped to sound with the same algorithm used in the camera based version.

Instead of supplying a rather complex GUI in which the user can create and customize presets (each one describing the relationship between motion and sounds) in this home version we include a set of predefined presets, easily accessible from a drop-down list. With this set we tried to cover a broad range of interaction modalities/gestures (arm motion, kicks, jumps etc.).

We included two special modes, designed for improving motor coordination. In the first one different movements trigger playbacks of a sampled voice pronouncing

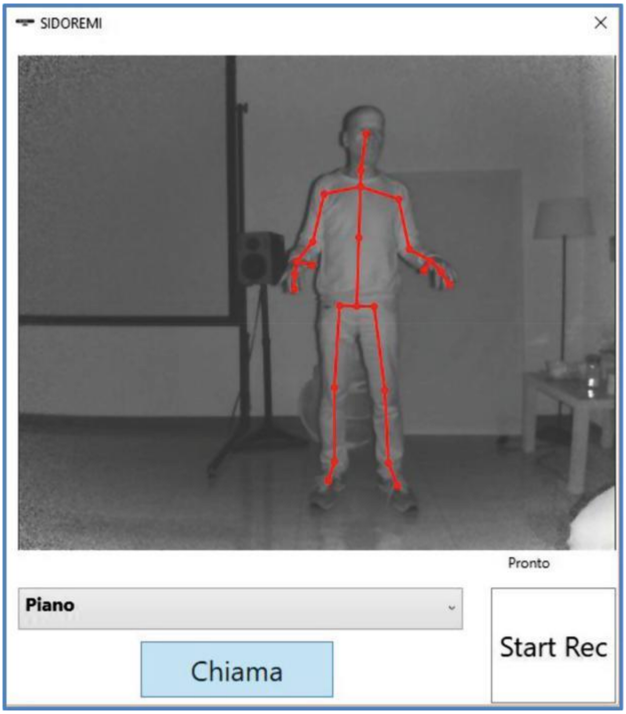


Fig. 3. Kinect based version GUI

numbers: the subject has to sequence movements according to numbers order (randomly shuffled at each program launch). The second one is similar but we use fragments of sentences instead of numbers: here the user has to sequence these fragments to reconstruct a story.

This home version is intended to be used by the children's parents, not necessarily skilled. For this reason, they could eventually need a remote assistance. In order to provide a simple way to provide assistance we included a video chat mechanism in the software: the user, simply pressing a "call" button can make a video call to our technical staff.

Movement Recordings. We included a "record" button: pressing it all the child movements are recorded on disk for off-line analysis (for diagnostic purposes). The psychologist team that support the project suggested to analyze movements of the subject during the various sessions in order to objectively evaluate the effects of systems. We basically extract these features, as a function of time:

Average of movements amount (whole body, upper, lower), Variance of movements amount (whole body, upper, lower), Average of movement speed (whole body, upper, lower), Variance of movements speed (whole body, upper, lower), Coordination (correlation between movements of different part of the body).

Since the concept of imitation is very important in the evaluation of autism spectrum disorder, we included an algorithm for computing cross-correlation of movements of two different bodies (the subject's one and the parent's one): this similarity index is strongly related to evidence of imitation (Fig. 4).

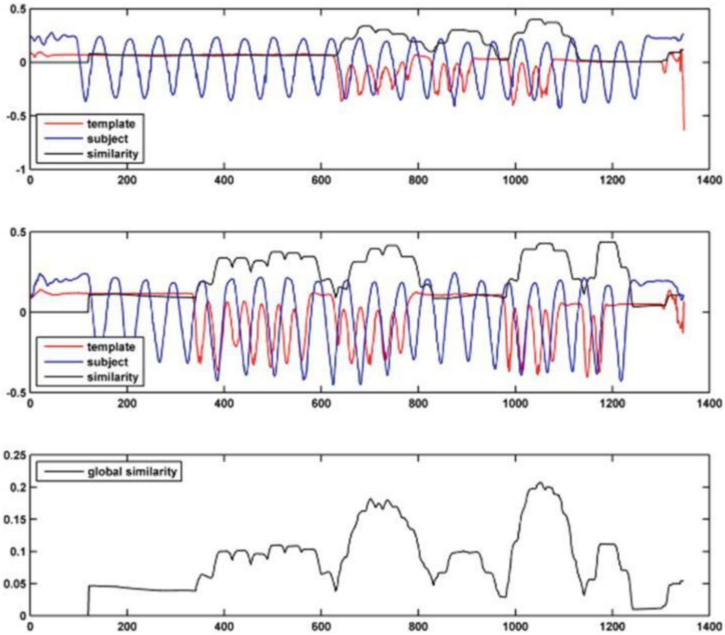


Fig. 4. Elbow movements in the two arms (upper graphs), for the subject and for his parent and (below) the resulting similarity index.

.

4 Experimentation

The experimentation [12] was performed in the school environment on 4 subjects (5–7 years, all males) diagnosed with Low-Functioning Autism (autism spectrum disorder, ASD). The weekly intervention lasted about 30 min. The children involved in the experimentation were evaluated in a cross sectional and follow up pilot study. Clinical features evaluated by mental health centers and information from the "Questionnaire on motor control and sensory elaboration" [13], compiled by parents, and from the "Short Sensory Profile" [14] filled up by the teachers were analyzed at baseline. Three clinical psychologists, not previously involved in the experimentation, analyzed the first eight videos of the intervention, completing an observation grid for every session. The grid was structured ad hoc by the research team on the basis of the DIR Floortime model and technique in relation to the benchmarks of the main sensory profiles. This grid was partially taken from the questionnaire of Politi and colleagues [15] aimed at assessing the sensitivity of music in children affected by autism spectrum disorder. The instrument is made up of nineteen items relating to the child's behavior during the sessions and measured the characteristics of each sensory profile in terms of the "four A": Arousal, Attention, Affection and Action [16].

In this experimentation to consider the sensory profile of the infant that undergoes the sound stimulation and interaction was crucial. The choice of sound stimuli related to the movement has been made individually for each child during the first sessions through broad-spectrum stimuli. All the interventions were calibrated on the basis of the observations drawn from the video of the previous meeting, viewed by the reference clinician, a child neuropsychiatrist. It is important to highlight that the clinical reference as well as the operator who conducted the interventions with children had formal training in DIR Floortime method (Fig. 5).

5 Results

The concordance rate between the three psychologists' behavioral observation grid was calculated with the interclass correlation coefficient. Moderate to good inter-rater agreement [interclass correlation coefficient (ICC) comprised between 0.596 (95 %CI 0.41–0.853) and 0.799 (95 %CI 0.489–0.933)] were found. A repeated measures design was performed to evaluate change over time for each child for the first eight sessions (T1–T8). The analysis of variance was performed to assess if there has been an improvement in specific symptomatic areas. The repeated measures analysis of vari-ance indicated an overall increase of the scores drawn up by psychologists (T1–T8; $p < 0.05$) (Fig. 6).

Concerning statistical indexes our study highlights that participants had improved several skills. These variations in behavioral expressions reflect a relational evolution indicating the beginning of an opening attempt to someone no longer perceived as a threat but as someone from which to draw contentment, through playful interaction with the sounds. This pilot study demonstrates positive results: children developed skills in establishing joint attention, imitation of caregivers, communicating with gesture and symbols (Fig. 7).

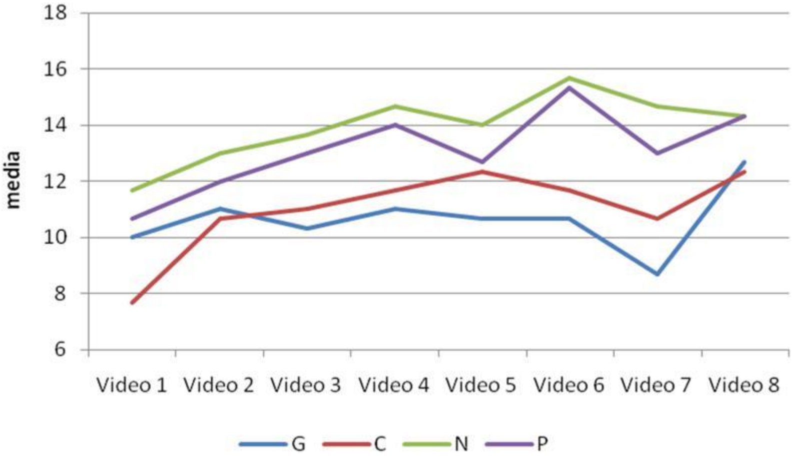


Fig. 5. Mean total scores of the behavior observation grid to assess change over time for each child (T1–T8).

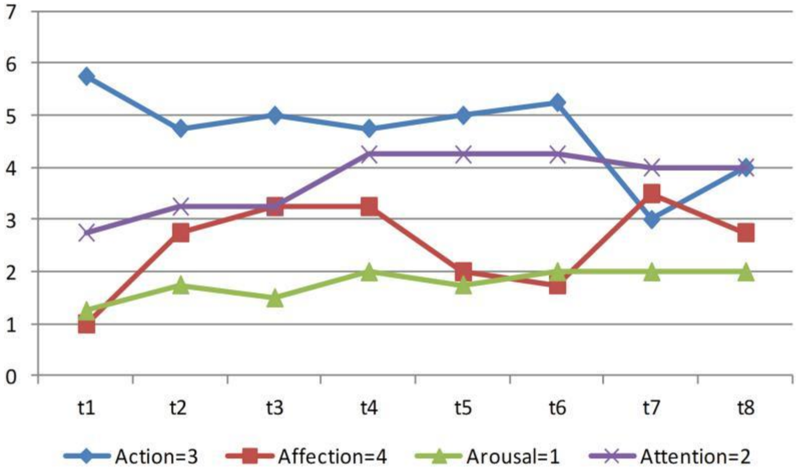


Fig. 6. Children's behavior (first eight videos) - Characteristics of each sensory profile in terms of Arousal, Attention, Affection and Action - From the observation grid of the first psychologist.

6 Further Development

The presented systems are audio feedback based, our goal is to provide multichannel feedback adding video interaction to the current system. We are testing different approaches, using Unity and Processing frameworks to enhance the visual experience. Three main demos are under testing: in the first one the subject movements are replicated by a 3D avatar (Fig. 8), both the avatar and the scenario can be customized and provide a great sense of immersion in an imaginary world. The second one is based

.

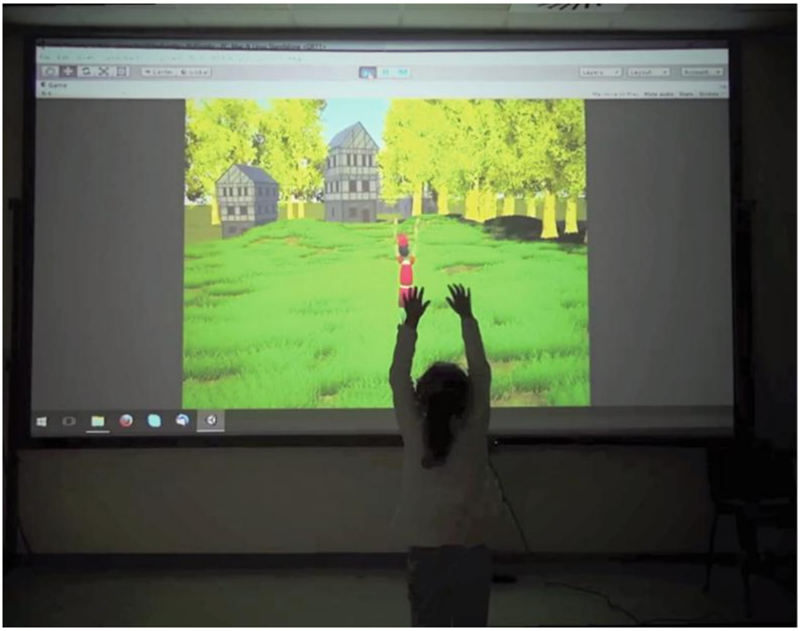


Fig. 7. Kinect controlled avatar

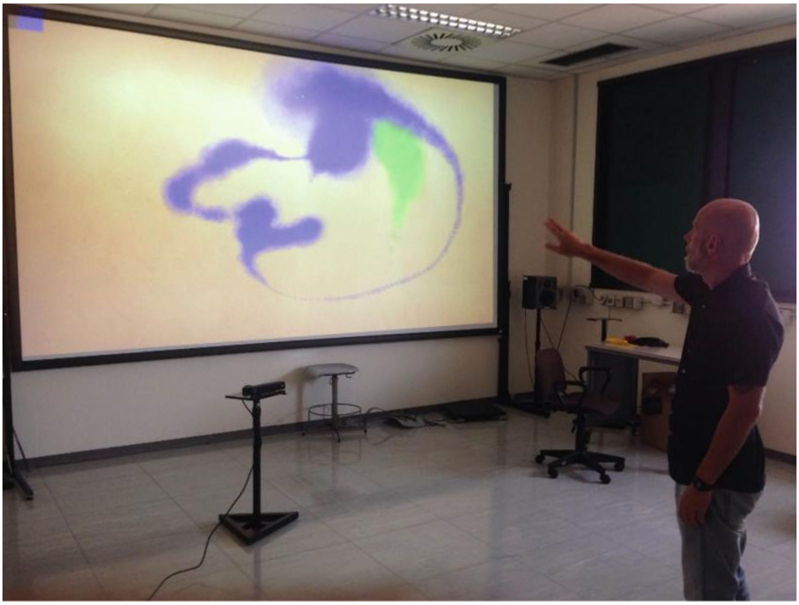


Fig. 8. Aerial painting system

on imitation, a video shows a trainer doing some move or postures and the subject have to correctly replicate the same movements. At the end of the exercise an algorithm assigns a score to the execution. The last demo is about aerial painting, the subject waves his hands in the air and those movements are mapped to brush strokes on a virtual canvas (Fig. 9). The next step of our work consists in the integration of these new activities in the current system and, in strict collaboration with the medical stuff, in the creation of new multichannel feedback based exercises to be tested in the next experimentation session.

7 Conclusions

We described an interactive, computer based system based on real-time image processing, which reacts to movements of a human body playing sounds. The map-ping between body motion and produced sounds is easily customizable with a graphical user interface. This system has been used for testing an innovative music-therapy technique for treating autistic children. The experimentation with real cases demonstrated several benefits from the application of the proposed system. These have been confirmed both by the team of clinical psychologists (using a validation protocol) and by the parents of the young patients. The most interesting outcome of the experimentation was the relational improvement. This promises to transfer the behavior shown in the setting to the external environment, increasing communication and interaction in the real world. We are continuing our experiments using Microsoft Kinect v.2. Our future approach will combine the Kinect's proprietary technology with the image processing techniques used in the camera based version.

The project addressed many challenges. Each autistic child is unique in the sense that improvements in abilities are very subjective and some study limitations have to be mentioned: first of all, the small size and the non-homogeneity of the sample; this is due to the difficulty of enrolling Low-Functioning Autism children with similar profiles. Participation is self-selected and sample bias cannot be excluded. A more rigorous assessment and selection of participants and the selection of a matched control group will guarantee results of higher value. Moreover, progress trends could depend upon external factors such as family involvement and health/treatment conditions. The outcome measurement also presented some limitations: our main measurement was the observational grid that is not a standardized instrument; moreover, information on important outcomes was not measured, such as cognitive skills and school perfor-mance. These data would be extremely interesting for creating the bases for using accessible technology-enhanced environments.

References

1. Magrini, M., Carboni, A., Salvetti, O., Curzio, O.: An auditory feedback based system for treating autism spectrum disorder. In: REHAB 2015 Proceedings of the 3rd 2015 Workshop on ICTs for Improving Patients Rehabilitation Research Techniques, pp. 30–33. ACM, New York (2015)
2. Ould Mohamed, A., Courbulay, V.: Attention analysis in interactive software for children with autism. In: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, Portland, Oregon, USA (2006)
3. Kozima, H., Nakagawa, C., Yasuda, Y.: Interactive robots for communication-care: a case-study in autism therapy. In: International IEEE Workshop on Robot and Human Interactive Communication (2005)
4. Villafuerte, L., Markova, M., Jorda, S.: Acquisition of social abilities through musical tangible user interface: children with autism spectrum condition and the reactable. In: Proceedings of CHI EA 2012–CHI 2012 Extended Abstracts on Human Factors in Computing Systems, pp. 745–760. ACM, New York (2012)
5. Riva, D., Bulgheroni, S., Zappella, M.: Neurobiology, Diagnosis & Treatment in Autism: An Update, John Libbey Eurotex (2013)
6. Pajareya, K., Nopmaneejumruslers, K.: A pilot randomized controlled trial of DIR/ Floortime™ parent training intervention for pre-school children with autistic spectrum disorders. Autism 15(5), 563–577 (2011)
7. Solomon, R., et al.: Pilot study of a parent training program for young children with autism The PLAY Project Home Consultation program. Autism 11(3), 205–224 (2007)
8. Wing, L., Gould, J.: Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. J. Autism Dev. Disord. 9, 11–29 (1979)
9. Vismara, L.A., Rogers, S.J.: behavioral treatments in autism spectrum disorder: what do we know? Annu. Rev. Clin. Psychol. 6, 447–468 (2010)
10. Kanner, L.: Autistic disturbances of affective contact. Nervous Child 2, 217–250 (1943)
11. Greenspan, S., Wieder, S.: The Child with Special Needs. Perseus. Pub., New York (1998)
12. Magrini, M., et al.: Progetto "SI RE MI" Sistema di Rieducazione Espressiva del Movimento e dell'Interazione. Autismo e disturbi dello sviluppo, Erickson (2015)
13. De Gangi, G., Berck, R.: DeGangi-Berck: Test of Sensory Integration. Western Psychological Services. Los Angeles (1983)
14. Dunn, W.: Sensory Profile-School Companion Manual. Psychological Corporation, San Antonio (2006)
15. Politi, P., Emanuele E. e Grassi, M.: The Invisible Orchestra Project. Development of the "Playing-in-Touch" (PiT) questionnaire. Neuroendocrinol. Lett. 33(5), 552–558 (2012)
16. Meini, C., Guiot, G., Maria Teresa Sindelar, M.T.: Autismo e musica. Il modello Floortime nei disturbi della comunicazione e della relazione, Erickson