Non-invasive attitude detection for full-body interaction in MEDIATE, a multisensory interactive environment for children with autism.

Narcís Parés, Anna Carreras, Miquel Soler.

Experimentation on Interactive Communication group (http://www.iua.upf.es/eic)
Audiovisual Institute, Universitat Pompeu Fabra (Barcelona, Spain)
Email: `npares@iua.upf.es`

Abstract

This paper presents the integration of simple well known vision techniques into a complex environment for children with severe autism. The interest lies in describing how this vision system has been designed such that it adapts to the strong constraints set by the environment. Specifically: (a) non invasive to the users, (b) cope with variable visible lighting and background, (c) full body interaction for the users, (d) real time detection, (e) defined ambiance, (f) adapt to a wide spectrum of non-typifiable users, (g) transportability (relatively light weight, fast set-up and small footprint), (h) reduce costs and (i) safe, robust and sturdy.

1 Introduction

MEDIATE (A Multisensory Environment Design for an Interface between Autistic and Typical Expressiveness) is an interactive environment that generates real time stimuli (visual, aural and vibro-tactile) such that low functioning children with autism, that have no verbal communication, can express themselves and dialogue with the environment creatively in a relaxed attitude. MEDIATE provides a controlled environment where the child has no external distractions and allows for a clear action-reaction full body interaction. This gives the children a strong sense of control and the notion that they can exert this control; i.e. what is called the *sense of agency* [1].

Autism is not a unique disability but rather a spectrum of disabilities, hence the design of an interactive system for children with severe autism is one of the most challenging areas for a designer. Specifically the user cannot be typified making standard design strategies inapplicable and non-invasive sensor systems must be designed to capture the children’s attitudes. The *brain* of the environment (i.e. the decision maker module and the pattern recognition module) must also adapt to each child by sensing repetitive attitudes (e.g. hand flapping, which usually means the user is not feeling at ease) and modulate stimuli generation such that the user is maintained in a calm state.

MEDIATE is a thirty months project funded by the EC under the FP5 / IST / Systems and Services for the Citizen / Persons with special needs (including the elderly and the disabled) [2].

In this paper we will describe the MEDIATE artificial vision capture system and how it is linked to the interaction system design and functionality. Although users interact in MEDIATE with visuals, sound and vibration, in this paper we will restrict the description to visual interaction for the sake of simplicity and also because the artificial vision system and the visual interaction system have both been designed and developed by our group.

2 Autism

Autism is a set of disorders in intercommunication and interrelation abilities that lead to an impairment of cognitive and emotional development. The essential characteristics of this disorder are the presence of an abnormal development in the following areas:

- Communication: Difficult or inexistent verbal communication. Difficulties in non-verbal communication.
- Socialization: Severe difficulties in interpersonal relationship.

- Imagination: There is a lack of imagination characterized by repetitive game play.

The factors that determine autism have a biological cause and the disorder is manifested during the first thirty months of the child. At a cognitive level, there is a *weak central coherence* [1] that impairs an adequate integration of the stimuli in the environment. This is externally translated in a lack of affective expression, an apparent lack of empathy, an obsessive concentration on particular elements and, often, repetitive movements. These three main characteristics make the child unable to discriminate and, more importantly, predict any of the events that occur in daily life. Hence, this unpredictable and overstimulating world is felt as alien making them feel isolated, i.e. they loose the sense of agency.

There are huge differences among individuals that are placed at different levels of the wide spectrum of autism. MEDIATE, is focused on children between 6 and 12 years of chronological age; low functioning persons in the autistic spectrum (PAS) with no verbal communication abilities.

Because the spectrum of disorders in autism is so wide, we had the imposed restriction of not being able to typify the user. Hence, we have had to establish new strategies in interaction design and develop new specific interfaces to make them non-invasive for these autistic children. This implies that no sensors or cables could be placed on the user. We could not even consider the child wearing markers nor dressing in any specific manner. Hence the sensors had to be all external and always compatible with the real time stimuli generation system.

3 MEDIATE environment

MEDIATE is an irregular hexagonal space, approximately six meters in diameter (for a full justification see [2] & [3]). Inside the space several elements act as interaction interfaces (Figure 1):

- Floor surface: it reacts to footsteps generating sound.
- The tune fork: a wall with tube-like structures that generate sound when caressed or stroked.
- The screen walls: two large (3 x 2.35 m) rear projection screens used as the support for visual interaction that react to the child's movement and touch. They allow the user to have a full body interaction with images in a 1:1 scale relationship.
- The impression wall: a wall with padded structures that react to pressure and emit vibration.
- The sound interface: a set of microphones and speakers that react to sounds emitted by the child in the space (voice, clapping, etc.).

These multiple modalities have been included to cover the wide spectrum of potential preferences of the users and to explore cross-modality.

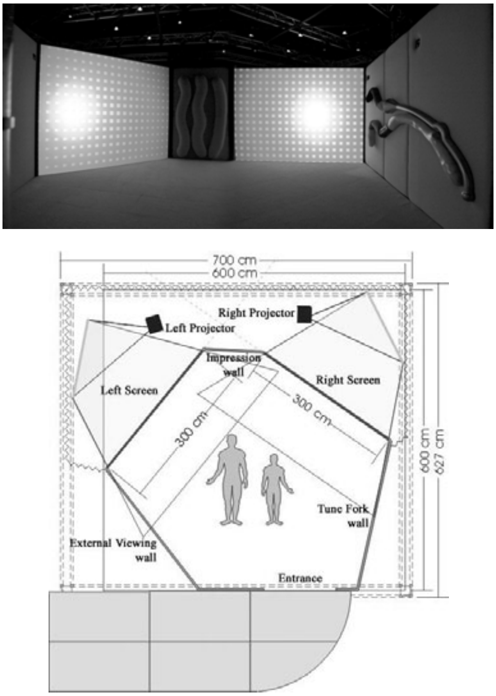

Figure 1. Environment: (a) panoramic view of the interior & (b) interaction elements.

As said before, we will concentrate on interaction with visual stimuli only, to explain the criteria used in designing the non-invasive artificial vision capture system.

4 Interaction with Visual Stimuli

We will first describe the interaction design criteria to then be able to describe and justify the design criteria of the vision system.

Interaction design

The common procedure in the design of interactive systems is to start by defining the type of user and application. Because we could not typify the user for MEDIATE, a new approach was advisable. We decided to apply a design strategy that our group had formalised in previous work: the interaction-driven design as opposed to a content-driven or user-driven approach [4][5][6]. Thus, we started by identifying the input/output interfaces, then defined an *interaction model* and finally defined the type of application and visual elements to be used. In other words, we started by concentrating on how the user is to *interact* with the application, analysing the interfaces, interaction with the elements, participation/ manipulation/ contribution of the user, in such a way that the obtained results allow a spontaneous emergence of the specific topic, content, “aroma” or tone of the application.

The project consortium had agreed that users would have to be interested and engaged in our environment because of the interaction itself. In other words, we wanted the children to gain a sense of control, because this would hopefully make them feel at ease and prepared for communication. Obviously we had to consider the fact that our users are children with autism that have no verbal communication capacities and that may have low resolution motor control.

Natural full body Interaction

Therefore, if the accent was to be placed on such interaction it had to be based on very clear action-reaction situations (also called *contingent* situations) and probably rely on full-body interaction.

Because of this and because we could not typify user behaviours for the type of users we were dealing with, we decided to find basic and very general body behaviours that could easily cause a reaction in the system and that would be clear to the user. These very simple behaviours that any child should be able to do were:

- move in relation to the screen,
- gesticulate in front of the screen,
- touch or lean on the screen.

By simply moving through the defined space, the environment already starts responding, opening small doors that can lead the child into playing with it. But also the other listed basic body behaviours are picked up by the vision system and used in the interaction dialogue. (Other sensing subsystems also detect pressing, screaming, clapping, etc., through specific sensors such as: microphones, pressure gauges and transducers.)

Visual Stimuli

As stated above, we wanted the children to be interested and engaged in MEDIATE because of the interaction itself. This implied that it would not be adequate for them to be engaged because of any of the possible contents within the environment. For example, if a dog appeared on the screens and a child had had a bad experience with dogs or on the contrary the child loved dogs, then the whole environment could be ignored by the child because of her concentration on that specific content and hence the whole experience would be a failure.

The result of this reasoning was that we decided to work with abstract or non-representational images.

In our search for visual design, we found that children with autism are very fast at finding a shape hidden in a mesh of lines. On the other hand, we also found that children with autism have difficulties integrating parts into complex objects and rather perceive groups of individual objects (*weak central coherence* [1]). Because of these two issues, we thought of working with isolated geometrical elements and this immediately reminded us of particle systems (see for example, [7]). Each particle could have its own particular behaviour; or the whole group of particles could have a group behaviour; or the group could have a global behaviour that came from the sum of individual behaviours. Also, particles could be individual isolated elements, or they could be grouped to form larger objects, or spread out forming a background, etc.

Interaction Models

With the idea of abstract, geometrical, contingent environments, we designed up to eleven interactive “games” with particles. Four were fully developed as preliminary work in the process of obtaining the final visual interaction for MEDIATE. These interactive “games” or “expressive experiences” we called *interaction models* because they not only set the rules of a game, but also define a philosophy behind the game that states what we are searching in the child’s play (for full description and justification see [2] [3]).

The final interaction model implemented inside MEDIATE environment is based on a screen tiled with square particles. Initially though, the screens are empty, only flooded with an initial colour that sets the interaction gamut. When the child enters the environment, the camera vision system detects her presence and presents a grid of small tiled particles (Figure 2a). If the child backs out and exits, the environment hides the tiled particles and returns to the idle state with empty coloured screens. This basic game of crossing the entrance line back and forth has been very noticeable to the children, many of whom have successfully discovered and enjoyed it according to the psychologists’ observations.

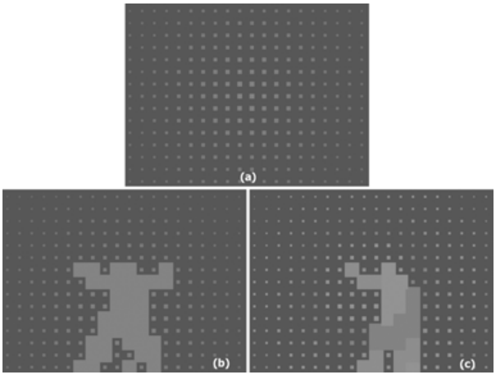

Figure 2. Final interaction model: (a) background particles showing radial gradient in size and shade, (b) a view of the gelatinous silhouette of the user, (c) a wave of colour being generated by the user touching the screen.

The particles grow as the user comes closer to the screen and shrink as the user moves away. There is a gradient in size and shade of colour of the particles from the user’s position to the edges of the screen, creating a constant sense of shelter wherever the child moves to (Figure 2a). When the child is in front of the screen, the particles that fall within the area of what would be her projected silhouette, grow to reveal a blocky silhouette (Figure 2b). The movement of the silhouette is done by interpolating the growing and shrinking of the particles, thus giving a sense of gelatinous material. This gives interaction a very fluid feeling. Finally, if the child comes very close to the screen and/or touches it, a wave of colour is generated starting from the touched point outwards (Figure 2c).

5 Capture system

Non invasive sensors

As stated before, an essential precondition in the design was the use of non-invasive sensors. Hence the sensors had to be all external, distributed throughout the environment and analysed and managed by the computer system in real time.

Regarding the requisite of enhancing a full-body interaction, we had to think of methods to sense the position, gestures, attitudes, displacements, etc., of the user in a non-invasive manner. The robust solution found for this was an artificial vision system based on cameras distributed throughout the environment that could communicate with the system to sense a set of properties of the user.

Artificial vision camera system

After analysing the required information on the user that the system had to sense, we came to the conclusion that it was not necessary to use a 3D motion tracking system.

On the one hand, we really only needed the position of the user with respect to the floor plane in order to have a rough idea of where the user was at each moment in time. So in this sense we only needed a 2D tracking system.

On the other hand, because we mainly needed the silhouette of the user to interact with its "reflection" on the screen, we only needed to read this silhouette with respect to two vertical planes, each one being parallel to each screen (Figure 1b). This is also a 2D vision system only. To be able to have silhouette readings almost throughout the environment we also defined two more vertical planes, each one opposite to each screen covering the silhouette when the user interacts with the "Tune Fork" and when the user is around the external viewing wall (Figure 1b).

As we have said previously, the *brain* of the system needs to detect whether the user is doing repetitive movements to try to modulate the stimuli generation. Therefore a rough idea of head and hand movement was also found to be useful at any moment in time. We thought that detecting 2D position of head and hands with respect to the four vertical planes defined for silhouette would allow for a good evaluation of repetitive movements.

Finally, we also required to detect when the user was touching (or extremely close to) each of the screens. As described later, we placed one camera behind each screen to track hard shadows (*blobs*) on them such that the effect was that of giant touch screens (Figure 3). This also defines a 2D vision system on each screen.

This fortunately ruled out 3D capture systems that would have had the following disadvantages:

- very difficult and laborious individual user calibration especially inadequate for children with autism and not very suitable for a transportable system.
- they need strong illumination and markers on the user so they are invasive, when we actually needed non invasive systems.
- they give very noisy output data that is not too useful for real time systems and is most often used off line so that it can be cleaned and smoothed.

This led us to design a system with 9 greyscale video cameras, some of them working together and some individually to capture the needed properties of the user. They were placed inside MEDIATE environment as shown in Figure 3. Cameras position and orientation was initially studied with a 3D virtual model.

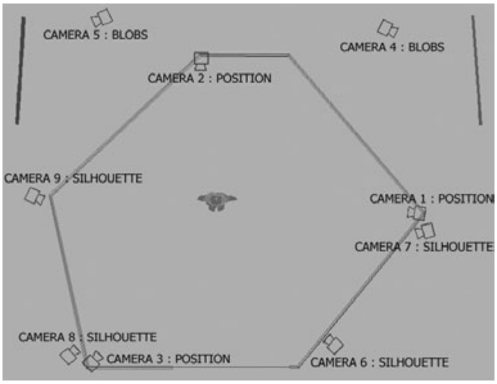

Figure 3. Capture system camera arrangement: position detection (cameras 1, 2 & 3), touch screens (4 & 5), silhouette detection (6, 7, 8 & 9).

Position is obtained by three cameras that jointly cover the whole of the floor space within the environment: cameras 1, 2 and 3 (in Figure 3). Each camera obtains the image point of the situation of the user inside the environment with respect to the floor plane (Figure 4). By applying the inverse perspective deformation of the camera the space point coordinates are obtained. This calculation is based on a DLT (direct linear transform), from each camera point of view. [8][9][10][11]. Then a correction of the calculated value is done by redundancy. A single point on the world floor is obtained through linear balancing of the three points from each camera view. Nevertheless, we do not always use the three points to obtain the final point because there are special cases in which the calculated point from one of the views cannot be considered as valid.

The user position detection cameras require a simple general calibration process used to obtain their projection matrix in order to calculate properly the inverse perspective deformation of the floor portion they see. The calibration of the three position cameras implies only acquiring the co-ordinates of four specific floor reference points. This means the user is not involved in this calibration process, which would otherwise be inadequate for autistic children. The accuracy of the position has been evaluated to be ±10cm in average, which in relation to the type of displacements to be detected is precise enough.

The three cameras that are used to detect the user position with respect to the floor plane are placed at approximately 235cm above the floor on top of the walls. The views of the position cameras is shown in Figure 4.

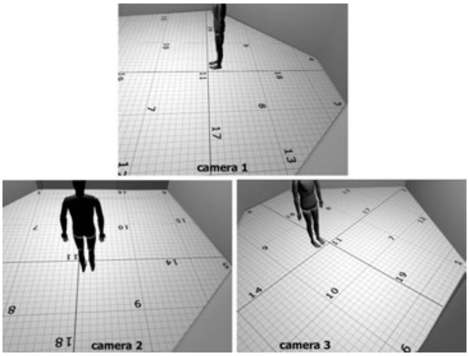

Figure 4. Camera capture simulation showing a grid for referencing views of position cameras: camera 1, 2 and 3 (Figure 3).

Two other cameras are used to detect user touching the screen (cameras 4 and 5 in Figure 3). Each one is placed behind a screen at a height of 115cm, such that their axis is perpendicular to the screen plane. Actually, as shown in Figure 3, the cameras see their corresponding screen through the mirror (in a similar way to how the projectors use the mirrors) such that the lens of the camera needed not be too wide angle but such that the whole footprint of the environment was as small as possible. The way this sensing is actually achieved, is by illuminating the screens smoothly and uniformly and then detecting hard shadows casted by the user (especially hands) on the screen that signify extreme proximity, touching or leaning. The precision of the touched point is ±1cm, although the minimum size of the detected blob is 10cm.

As described before, due to the required sensing and to the space distribution, four cameras are set to appropriately sense and track the user silhouette and her 2D skeleton in the four defined vertical planes: the right and left screens (cameras 8 and 6 respectively); the Tune Fork wall plane, actually including a part of the entrance to the environment (camera 9); the external vision wall plane, also including part of the entrance to the environment, (camera 7) (see Figures 1b and 3 for reference).

All these four cameras are hidden behind the environment walls and peek into the environment through especially defined holes in the walls to avoid the children from being distracted. They are at a height of 200cm such that they have a view of the silhouette of the user which is close to perpendicular to the vertical plane, but out of reach from the users.

Camera adjusting

In order to facilitate build up of the environment at each new site, a simple system that helps to correctly adjust the position and orientation of each camera has been developed. It is based on the comparison of what the camera should be seeing and what it is actually seeing at that set up moment.

The adjustment software subtracts a reference image from the current image that the camera is capturing in real time. If the current position and orientation of the camera is correct, the software yields a completely black image (perfect subtraction). If on the contrary, the obtained subtraction image shows white areas, it means that there are areas in the real time captured images that do not match the reference image. In this case, the operator must readjust the camera (Figure 5). This method has proved extremely fast and easy to use and the manual adjustment of the cameras can be done by two operators (one adjusting the camera and another checking the subtraction result) in about 3 minutes for each camera.

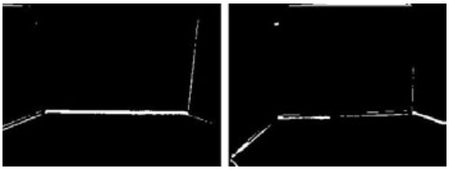

Figure 5. The subtracted images of what cameras 6 & 7 should be seeing and what they are actually seeing. The white areas show the misalignment that still exists in camera orientation and position.

Lighting

The artificial vision system had to deal with lighting the space and objects that will participate in the capture. The more controlled the lighting of the environment, the easier the capture and image processing to obtain the desired properties and data. The reason behind this being that basic operations like background subtraction, that help enormously in fast object segmentation, can be done straight away.

But of course in an environment like MEDIATE, with two large projection screens that give a variable background and a variable lighting intensity, the capture system can become extremely difficult to define and implement. Because of this we had to find a method to control lighting and not be affected by the aforementioned constraints.

Near Infrared Spectrum

The solution was to use near Infrared (IR) lighting and capture at around 850nm wavelength. Our research led us to discover that the projections of the screens only work in the visible range of light spectrum; i.e. they generate no frequencies within the near IR range and hence the cameras see them as if they were “off”. So we were able to make compatible the visual stimuli generation and capture systems to obtain the desired data for the visual *interaction model*. In other words, the visuals do not interfere in the capture process and the correct lighting of the environment and user does not limit the correct visualisation of the images on the screens by the user.

Using filters on the cameras to block out all visible range frequencies and letting through only the near IR frequencies, we obtain a completely uniform and stable environment that allows us to segment the user from the environment using a straightforward background subtraction process. The subtraction is an absolute value difference between each captured image and a reference image taken from the empty environment at the beginning of each session. This is only feasible because the background stays stable throughout the session for the IR band pass filtered cameras.

Under this set up, we then designed an adequate near IR lighting system for the environment and the user, such that we obtained clear and crisp images for the image processing algorithms and was almost invisible for users.

The idea was to light uniformly the whole space such that the user would always be correctly seen by the camera system. IR lights commercially available for security systems are expensive and because of the geometry of the space, we would have needed quite a few to smoothly light it all.

Halogen lamps generate a lot of intensity in the near IR range. Therefore, regular dychroic lamps were found to be a very cheap and adequate system to light the space. Nevertheless, they also produce very intense light in the visible range and therefore would disturb the correct viewing of the images on the screens, washing them out, and would not allow for a dim, cosy, carefully lit environment. To solve this, we adapted pyrex glass IR filters that allow only a band around 850nm to pass through. This way the environment is not affected by visible light and the user is still clearly seen by the cameras.

The final arrangement consists of a grid of 26 dychroic halogen lamps with an IR filter each, approximately 80cm apart from each other and from the walls, at a height of 235cm from the floor (Figure 6).

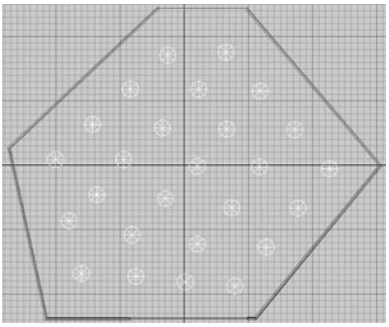

Figure 6. IR MEDIATE lighting grid.

This arrangement also allows a smooth lighting of the screens to be able to use them as giant “touch screens”.

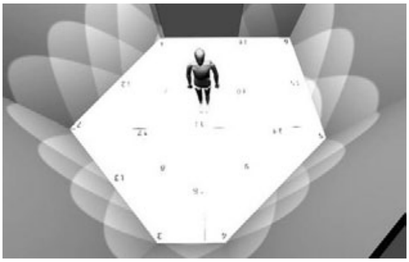

Figure 7. IR MEDIATE lighting intensity simulation. Interior view from the top opposite side from the space entrance.

Image capture

In order to capture images coming from 9 cameras simultaneously and in real time (25 fps), we used three RGB FlashBus capture boards. Each board has three greyscale cameras connected to it as if the three greyscale images were in fact the three channels of a single RGB image. The three greyscale cameras are synchronised from the board to guarantee that their images reach the board at the same time.

Two of the boards were installed in one powerful PC (Intel-based), thus dealing with 6 cameras at the same time, while the third board was installed on another similar PC. This allowed us to minimise both capture hardware and PCs needed to process the images and generate sensed data.

The images are captured at a resolution of 320x240pixels with 8 bits per pixel and at 25 frames per second.

Image processing

To process the greyscale images we decided to use a tool called EyesWeb. This software is the result of another EC funded project led by the University of Genoa [12][13]. EyesWeb is fully based on the Intel Open CV vision libraries, but it has a visual programming interface that allowed us to quickly, easily and efficiently prototype all our image processing algorithms through block patches composed of small simple atomic image processing operations.

Briefly, all the patches start by capturing the RGB image from one capture board, separating each RGB channel to process it separately (as a single greyscale image) and applying background subtraction to segment the user as described in above.

From this point each patch applies the relevant processing to the segmented user image in order to obtain the desired data described in previous sections.

The data is finally sent via TCP/IP to an intermediary application that we implemented. This application does the position calculation of the user, gathers the rest of information calculated within EyesWeb, formats it according to MEDIATE's messaging system and sends it via TCP/IP to the rest of the system.

6 Results

The artificial vision system designed and developed obtains all the data needed for the planned interaction. It allows the children to naturally enter the environment in a completely non-invasive system that does not overwhelm them, does not limit their mobility and does not bias their attitudes.

MEDIATE is a transportable environment and sessions with children with autism have been held in London (Goldsmith's Institute, Kings College) during two weeks, in Hilversum (Hogeschool voor de Kunsten Utrecht, Netherlands) for five weeks, in Barcelona (Universitat Pompeu Fabra) for four weeks and in Portsmouth during six more weeks.

While in Barcelona the eleven children that have participated in MEDIATE have passed three times each. Hence thirty-three, two hour, sessions have taken place here.

The results extracted from those experienced sessions with autistic children are that only one girl did not want to enter the environment on her first visit. On her second visit she had no problem to enter and play. The rest of children played in greater or lesser degree, but they entered without effort. Needless to say that the children are not forced to enter in any way. In fact, the psychologists ask the parents (who are present during interaction just outside the action space) not to push (neither physically nor verbally) their children to enter the environment. This is already a huge success for the environment and the lighting and capture vision system.

These children, that need very rigid daily routines and that do not cope well with unknown places, have actually become curious enough to enter by their own will and start playing. The time spent in the environment has varied from 5 minutes to 35 minutes. In every case, it has been clear that the children have found at least one of the proposed visual interactions and have successfully played with it. For example many children found the simple and clear game to make the particles appear and disappear on entering and exiting the environment, playing at the entrance threshold, which greatly help them in gaining a sense of control.

According to the psychological results, none of the children felt uneasy or uncomfortable in the environment (only one of the sessions had to be stopped because of overexcitement of the child) and most appeared to gain the desired sense of control and agency. Feedback from parents and carers of children who had used MEDIATE, shows they felt it was a hugely beneficial experience that they would like to be able to continue to use. More specifically it highlighted three main areas of benefit, these being: (1) Independence, (2) Person-centred and (3) No parental demands. In other words, the MEDIATE experience allowed many children for the first time to be completely on their own and be safe to make their own choices, enjoy their behaviours and get some recognition from the environment as to what they were doing. Moreover, this was achieved taking careful consideration of their particular sensory needs and communication difficulties. Finally, in the MEDIATE environment, many of the children could do as they pleased without having to meet the expectations or demands of others, including their parents or carers.

References

[1] Happé, F. *“Autism, an introduction to psychological theory”*, Psychology Press Ltd, Taylor & Francis Group, 1999. Leslie Lamport.

[2] Official MEDIATE project web site: http://web.port.ac.uk/mediate/

[3] Parés, N, et al., *"MEDIATE: An interactive multisensory environment for children with severe autism and no verbal communication."*, in Proceedings of International Workshop on Virtual Rehabilitation 2004 (IWVR 2004), Lausanne, Switzerland, September 2004. (Accepted Paper)

[4] Parés, N.; Parés, R. *"An Interaction-driven Strategy for Virtual Reality Applications"*. In: Abstract Proceedings of the VR World Congress, El.pub, IST, EC. Barcelona: www.VREfresh.com, 2001.

[5] Parés, N.; Parés, R. *"Interaction-driven virtual reality application design. A particular case: 'El Ball del Fanalet or Lightpools"*. In: PRESENCE: Teleoperators and Virtual Environments. Cambridge, MA: MIT Press, 2001. Vol 10.2. Pag. 236-245. http://www.iua.upf.es/~npares/publicacions/interaction-driven.pdf

[6] Parés, N; Parés, R. *“Una estratègia basada en la interacció per a aplicacions de realitat virtual”*. Paper at CAiiA-STAR Symposium, Barcelona 2001. http://www.uoc.edu/artnodes/cat/art/ilustrat_pares0902/ilustrat_pares0902.html.

[7] Foley, J. et al. *“Computer Graphics Principles and Practice”*. 2nd. ed., Reading, MA: Addison-Wesley, 1990.

[8] Hartley, R.; Zisserman, A. *“Multiple View Geometry in computer vision”*. Cambridge University Press.

[9] Faugeras, O. *“Three-dimensional computer vision : a geometric viewpoint”*. Cambridge, MA: MIT Press, 1994.

[10] Rogers, D.F.; Adams, J.A. *“Mathematical Elements for Computer Graphics”*. Ed McGraw-Hill.

[11] Hill, F.S. *“Computer Graphics”*. Ed Maxwell MacMillan.

[12] Official EyesWeb project web page: http://www.eyesweb.org/

[13] A. Camurri, M. Ricchetti, R. Trocca. *“EyesWeb - toward gesture and affect recognition in dance/music interactive systems”*, in Proc. IEEE Multimedia Systems '99, Firenze, Italy, June 1999.