Hindawi Publishing Corporation
Journal of Sensors
Volume 2013, Article ID 564864, 12 pages
http://dx.doi.org/10.1155/2013/564864

Research Article

A Multimodal Learning System for Individuals with Sensorial, Neuropsychological, and Relational Impairments

Sergio Canazza[1] and Gian Luca Foresti[2]

[1] *Department of Information Engineering, University of Padova, Via Gradenigo 6/A, 35131 Padova, Italy*
[2] *Department of Mathematics and Computer Science, University of Udine, Via delle Scienze 206, 33100 Udine, Italy*

Correspondence should be addressed to Sergio Canazza; canazza@dei.unipd.it

Received 1 February 2013; Revised 20 June 2013; Accepted 20 June 2013

Academic Editor: Ignacio Matias

This paper presents a system for an interactive multimodal environment able (i) to train the listening comprehension in various populations of pupils, both Italian and immigrants, having different disabilities and (ii) to assess speech production and discrimination. The proposed system is the result of a research project focused on pupils with sensorial, neuropsychological, and relational impairments. The project involves innovative technological systems that the users (speech terabits psychologists and preprimary and primary schools teachers) could adopt for training and assessment of language and speech. Because the system is used in a real scenario (the Italian schools are often affected by poor funding for education and teachers without informatics skills), the guidelines adopted are low-cost technology; usability; customizable system; robustness.

1. Introduction

Learning systems providing user interaction within physical spaces have been carried out over the years. However the high cost and the high complexity of the technologies used have always implied that their use by pupils in real context was limited to occasional visits or short periods of experimentation. Our aim is to provide an interactive multimodal environment (developed in C++) that can be integrated with the ordinary educational activities within the school. For this purpose, we use common technologies—such as webcams, microphones, and Microsoft Kinect sensors—in order (i) to provide tools that allow teachers to adapt or create autonomously the educational activities content to be carried out with the system and (ii) to implement a user interface for the management software that does not require specific computer skills.

Our system implements the five different types of interaction stated by Moreno and Mayer [1]: (1) dialogue, (2) control, (3) manipulation, (4) search, and (5) navigation.

Indeed, these five levels are very familiar during the everyday learning activity. Here are some examples: (1) a comparison/oral discussion in which the exchange of information is not unilateral, but the opportunities to the students to ask questions and express their opinions are given influencing the content of the lesson; (2) oral exposure in which the student has the ability to control the speed and to stop the explanation in order to benefit from the educational content at their own pace; (3) a scientific experiment that leaves the possibility for the student to test different parameters and see what happens; (4) the ability to independently seek information on a certain subject within a collection; (5) the ability to customize the use of the educational content through multiple paths, hypertext, and so forth.

A hardware and software platform and its validation are presented. The system consists of a set of tools for the improvement of listening comprehension in various populations of pupils having different disabilities, sensorial (deafness), neuropsychological (specific language impairment), genetic (Down syndrome), and relational (autism), and for the assessment of speech perception and discrimination, both in Italian and in immigrant pupils who learn Italian as their second language.

The guidelines adopted for the system design are as follows.

(1) Low-cost technology (computers, sensors, interfaces, etc.): it is necessary because the system has to be adoptable in schools of all levels, regardless of economic differences.

(2) Usability of the system: with the aim to obtain the widest distribution of the system, it is essential that teachers (that often do not have advanced skills in computer science) are able to use/control the system.

(3) Customizable system: the teachers have to be able to adapt the content of educational activities carried out in the system.

(4) Robustness of the system: the system must operate in a wide range of conditions, regardless of the environment in which it is installed. This requirement is necessary to reduce the need for the school to request technicians for system maintenance, in order to minimize additional investment.

The system conveys multimedia contents in the physical space, which can be a classroom, and it has the following main components.

(i) *Input components*, sensors (webcams, Microsoft Kinect, and microphones) for tracking the full-body movements, gestures, speech, and sounds.

(ii) *Real-time processing components*, a control unit (hardware and software) for the processing and mapping of the user's actions/sounds.

(iii) *Multimedia output components*, such as loudspeakers, headphones, and projectors to reproduce audio/visual contents.

The installation in physical space allows users to interact with this multimedia content in real time. The scope of designing and developing multimodal systems for learning is based on the assumption that the educational context needs to be renewed through the introduction of new educational tools and methodologies able to implement a high level of interactivity and multimodality. The need for a methodological renewal is due to the fact that, nowadays, teaching occurs in a highly heterogeneous educational context: there is a significant diversification in levels of learning, a high proportion of foreign pupils, and a growing number of pupils with disabilities. This heterogeneity entails a major transformation in the way our schools are organized. It suggests that teachers need to be trained to present their lessons in a wide variety of ways using music, cooperative learning, art activities, role play, multimedia, and technologies, in order to recognize and nurture all human minds and their combinations, to encourage interaction with the world, the general development of the person, and the achievement of the highest possible level of learning [1, 2].

Today, information and communication technologies (ICTs) are recognized as useful instructional tools [2]. ICTs indeed support teachers by offering a variety of computer-based learning activities by using computers, interactive whiteboards, and wireless tablets which can supply pliability and creativity in the processes of learning/teaching. Nevertheless, these ICTs lack the involvement of a physical environment, which is essential in education since a critical part of the pupils' cognitive development is their interacting with the physical world, within a social and cultural environment [3].

The main idea is to bring *physicality* into *brains-on* learning, combining full-body movements and gestures within a broad physical space with higher cognitive activities like thinking, reasoning, and reflecting [4]. This approach is based on a fundamental developmental assumption: effective learning takes place when meaning is taken from experience with the things of the world [3, 5].

The work answers specific problems of the European education system:

(i) how to manage an extremely heterogeneous school environment (due to mixed abilities, the growing number of immigrants and pupils with various disabilities);

(ii) how to actualize an effective teaching of foreign languages;

(iii) how to compensate the lack of physical activity in pupils' life; according to the European Commission for the Health Sector, this is public health problem which requires the implementation of specific actions to resolve it.

Our system implements two language functions: *higher level function*, namely, listening *story comprehension* which involves both cognitive and linguistic processes [6], and *lower level language* function, namely, speech perception and speech production [7].

A validation experimental protocol is developed and applied to a statistically significant number of samples (schools and medical institutions). The experimentation collected evaluative data about the system, with the aim of using it (i) to modify and improve the functionalities of the system and (ii) to assess the impact of the developed technological tools on the user.

The paper is organized as follow. Sections 2 and 3 provide the background, respectively, of audio and video features extraction; Section 4 describes the system architecture, the platform (Section 4.1) and the applications developed (Section 4.2). Finally Section 5 discusses the results and proposes some future directions of the system.

2. Voice Analysis: Features Extraction

Human voice can be characterized by several features, and human beings are able to control vocal emission in order to modify most of these features. Starting from the state of the art in the field of speech features extraction (see [7] for a review), we identified some prominent features, which were relatively straightforward to be extracted with real-time algorithms. They are as follows:

(i) intensity, computed as the root mean square (RMS), the square root of the arithmetic mean (average) of the voice pressure signal:

$$x_{\text{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}, \quad (1)$$

where x_i are the values of the voice pressure discrete signal;

(ii) spectral centroid: it is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the "center of gravity" using the Fourier transform frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes, or

$$\text{Spectral Centroid} = \frac{\sum_{k-1}^{N} kF[k]}{\sum_{k-1}^{N} F[k]}, \quad (2)$$

where $F[k]$ is the amplitude corresponding to bin k in discrete Fourier transform (DFT) spectrum;

(iii) a voiced/unvoiced flag, depending on whether the utterance is associated with pseudo-periodic vocal fold vibrations or not;

(iv) *pitch*, that is, the subjective attribute of frequency.

Although the first two features are defined as usual and easy to extract, the other two are not so trivially estimated, and therefore they are briefly described here. The voicing flag indicates the presence of a voiced signal, that is, a signal containing periodicities due to vocal fold vibrations. This kind of signals can be detected with various approaches. A technique which combines zero-crossing detection and cepstrum extraction [8] has been developed. Pitch estimation is facilitated in this case by the relatively simple harmonic structure of voiced utterances, so that the problem reduces to estimation of the fundamental frequency. This is estimated using an algorithm that extracts and matches harmonic spectral components on successive frames of the vocal signal [9].

The main idea was to construct a mapping of vocal features into well-recognizable graphic features.

3. Video Analysis: Tracking System

Interpretation of human gestures by a computer is used for human-machine interaction in the area of computer vision since the beginning of the eighties.

An exhaustive review of the state of the art and the currently available literature and techniques related to human body tracking is given in [10]. Some techniques are used to detect the body movement in 2D. Other segmentation techniques are designed to be used in bone and joint detection. Many schemes are proposed for tracking the body skeleton in 3D [11]. The common aim of all these techniques is to develop an automated tracking body or skeleton movement which can help create a digital character animation in 3D or control through some devices such as TV, robot, and so forth.

The method proposed by Akita's [11] recognizes the parts of the body through successive procedures beginning from the identification of the parts considered most stable and from a sequence of representative poses that the model can assume when carrying out a movement.

Some systems have tried to track human motion working on 3D models of the human body. In [12] O'Rourke and Badler's deduce the position of the various parts of the body by mapping the input frames in a 3D cylindrical model of the human body. The method proposed by Bregler and Malik [13] detects the 3D position of the parts of the user's body but requires to mark several segments of the body in the initial frame. Horprasert et al. [14] move a humanoid in a 3D virtual space by estimating the positions of the joints of the human body: the system is required to capture simultaneously several frames by means of several cameras located in different positions of the observed environment.

Some research used stereo cameras to estimate human poses or perform human tracking [15]. In the past few years, a part of the research has focused on the use of time-of-flight range cameras (TOF). Many algorithms have been proposed to address the problem of pose estimation and motion capture from range images [16]. Ganapathi et al. [17] present a filtering algorithm to track human poses using a stream of depth images captured by a TOF camera. Plagemann et al. [18] use a novel interest point detector to solve the problem of detection and identifying body parts in depth images.

The most popular tracking systems for human body pose estimation try to extract useful information either low level, such as edges, or high level, such as hands and head.

An example of low-level tracking can be found in the Walker system by Hogg [19], where edges are extracted from the image and matched against the edges of a human model to determine the pose of the human model.

There are several high-level tracking systems which try to track and/or estimate in real-time the 2D positions of the joints of the human body. They use statistical models of the background and foreground to segment the image into blobs [20]. Two examples of well-known methods are represented by the *Pfinder* [21] and W^4 [22] systems. *Pfinder* is a system for tracking people and uses a multiclass statistical model of color and shape to segment the subject from the background and to detect the 2D positions of the head and hands in a wide range of viewing conditions. Our system, unlike *Pfinder*, does not require a dynamic/kinematic model of the human body so it does not require very powerful computation systems. Furthermore, it uses color models (RGB and HSV) which are different from the YUV format used by *Pfinder*. W^4 is a real-time system for tracking people and for monitoring their behavior in an outdoor environment. It employs a combination of shape analysis and tracking to locate people and their parts (head, hands, feet, and torso) and to create models of people's appearance so that they can be tracked through interactions such as occlusions.

Most of the earlier techniques on gesture recognition were based on separate models of human body parts that are trained for each gesture, for example, hidden Markov

models (HMMs) [23]. Recent advances in gesture recognition research suggest that multiclass (one model for all gesture classes) models like hidden conditional random fields (HCRFs) [24]. HCRF provides an excellent framework for modeling each gesture as a combination of subgestures by sharing hidden states among multiple gesture classes. HCRF training algorithms are computationally very expensive as compared to those of HMMs.

From the body tracking survey, it is evident that the detection of the bone joints of human movement is still a major problem because the depth of the human body cannot be determined through the use of an optical camera. However, we have tried to use more than one video camera to detect and determine the depth of the object, but the consequence is that the cost increases and the process ability have been slowed down due to the increased data processing. Recently, a new impulse to the research in the field of real-time human action recognition has been given by the availability of an infrared or depth camera: the Kinect sensor. Xia et al. present a novel model-based method for human detection from depth images [25]. The method detects people using depth information obtained by Kinect in indoor environments. In particular, the system detects people using a two-stage head detection process, which includes a 2D edge detector and a 3D shape detector to utilize both the edge information and the relational depth change information in the depth image.

Our proposed system was conceived as a system for controlling a humanoid in a virtual world through the interpretation of user's movements [21, 26–28]. Since it is reasonable to assume that the subject keeps his legs relatively still in front of the monitor where the world is displayed, the capture process is limited to the upper section of the user's body. However, the system could easily be extended to track the whole human body, if needed. Furthermore, it requires the user to begin capturing in an erect posture with his hands placed away from the body and below the head; in this phase, several pieces of control information are used which are necessary to initialize the system's modules.

The silhouette is extracted by a change detection module of the system which uses a new algorithm for the extraction of moving objects from an almost static and arbitrarily complex background. The proposed algorithm is efficient and robust, and, unlike many existing systems, it is capable of detecting and eliminating from the scene the shadows produced by the movement of the user in the environment.

4. The System Architecture

The system constitutes a framework hierarchically structured in three levels (Figure 1).

(i) *Platform*. Hardware modules (camera or other capture devices, processing unit, speakers, projector, etc.) and software (input data processing, audio/video devices control, etc.) that provide the necessary infrastructure to make the system operative.

(ii) *Applications*. Software packages compatible with the platform, which implement the different interactive activities that can be performed with the system.

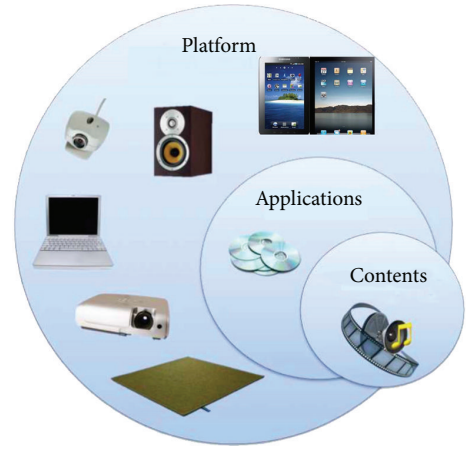


Figure 1: The logical architecture of the system, hierarchically structured in three levels.

(iii) *Contents*. Multimedia learning units broken down by subject area and designed for each application of the system.

4.1. The Platform. The platform is developed in C++ using openFrameworks, a toolkit for *creative coding* designed to assist the creative process by providing a simple and intuitive framework for experimentation. The toolkit is designed to work as a general purpose glue and wraps together several commonly used libraries. In our project we used OpenGL (for graphics); rtAudio, OpenAL, and FMOD (for audio I/O and analysis); Quicktime (for video playback); and OpenCV (for computer vision). In particular, open source computer vision (OpenCV) is a library of programming functions for real-time computer vision, released under a BSD license, and it is free for both academic and commercial uses (in line with point (1) of our guidelines, listed previously). It has C++ interfaces running on Windows, Linux, Android, and Mac. Because we used only the libraries of the toolkit core (without adding external parts, which can be dependent on the operating system), the code is cross-compatible. Our system supports five different operating systems (Windows, OSX, Linux, iOS, and Android).

Low-cost hardware is used: a personal computer with Windows XP, an MS Kinect sensor, a webcam, an infrared lamp, and a tablet (Android or iOS) where a dedicated app (linked to the platform) allows the teacher to control the behavior of the pupils.

With regard to the part of the software related to the video data processing, the principle aim of the system is to identify the user(s) and to track their movement. The motion-capture subsystem worked in three steps.

(i) Background learning/subtraction: acquired information on the background and subsequent isolation of the object of interest from the scene.

(ii) Optimization: noise removal and figure optimization by means of filters and morphological transformations. This phase is dedicated to the extraction of the

silhouette of the people in the room from the scene as accurate as possible.

(iii) Contour finder/blob tracking: identification and tracking of the moving objects by tracking the centroid of the blobs.

4.1.1. Background Learning/Subtraction. We adopted conventional techniques, based on the assumption that the adjacent pixels in the images are mutually independent. In some cases, this assumption is not applicable, and there are methods that for each pixel store information on its value and on the value of the surrounding pixels. These techniques require (at least) twice the memory and computational resources [29], and they are in contrast with the point (1) of our guidelines, listed previously.

The method used is based on the calculation of the background statistical modeling for each color channel and on the use of an adaptive threshold. After obtaining a statistical background model, $B(x, y, t)$, the new frame $I(x, y, t)$ is compared with it and the software produces a *confiance image*: that is, a gray-scale image where the value of each pixel represents the probability that it is part of an image region that does not belong to the background.

The background model construction for each channel is carried out calculating the mean and standard deviation of each pixel for a certain number of frames, acquired by static shots on the background:

$$\mu_t = \alpha x_t + (1 - \alpha)\,\mu_{t-1}, \tag{3}$$

where μ_t is the average calculated up to frame t, α is the learning rate of the model x_t, and the pixel value of the frame is t. This calculates a weighted average of exponential type of all previous values of the pixel. Subtracting the video frame after the acquisition of the background allow identifying which pixels have changed value.

Let us define standard deviation, for each pixel calculated as

$$\sigma_t^2 = \alpha\left(x_t - \mu_t\right)^2 + (1 - \alpha)\,\sigma_{t-1}^2. \tag{4}$$

The resulting images will be used for the normalization of *confiance image*. After acquiring the background model, successive video frames are subtracted from the average image, channel by channel. For each difference image, each pixel is normalized, using two thresholds $m\sigma$ and $M\sigma$ obtained from the standard deviation images. If the difference value is less than $m\sigma$, the *confiance* C^c (or the probability that the pixel analyzed corresponds to a region which is not part of the background) is set at 0%, and if it is greater than the $M\sigma$, *confiance* is set at 100%; if it is in an intermediate value, the *confiance* value is scaled linearly according to the expression:

$$C^c = \frac{D - m\sigma}{M\sigma - m\sigma} \times 100, \tag{5}$$

where D is the difference value. Since a variation in each color channel can indicate the presence of a region not belonging to the background, the final *confiance* image will have for each pixel, the maximum value among the corresponding pixels in the three color channels As regards the values of $m\sigma$ and $M\sigma$, they are obtained by the product between the image of the standard deviation σ and two scalars m and M.

Figure 2: Example of a classroom where our system is installed. The three windows do not have rolling shutters (and sun light on the floor is evident). The light conditions, of course, change from day to day, depending on weather condition, season, and time.

4.1.2. Optimization. The platform must operate in the real world (i.e., school rooms), regardless of the ambient conditions of the environment in which it is installed. Often the light is not controllable (e.g., windows without rolling shutters; see Figure 2). For this reason, the system uses signal processing procedures in order to make the platform particularly robust.

(i) Median filter (able to perform noise reduction on the image: see in Figure 3 an example of *confiance image* after applying a median filter).

(ii) Morphological operators [30, 31]:

(a) dilation: the basic effect of dilation on binary images is to enlarge the areas of foreground pixels (i.e., white pixels) at their borders. The areas of foreground pixels thus grow in size, while the background "holes" within them shrink;

(b) erosion: the basic effect of erosion operator on a binary image is to erode away the boundaries of foreground pixels (usually the white pixels). Thus, areas of foreground pixels shrink in size, and "holes" within those areas become larger;

(c) opening: it is a composite operator, constructed from the two basic operators described previously. Opening of set A by set B is achieved by first eroding set A by B, then dilating the resulting set by B;

(d) closing: like opening, it is also a composite operator. The closing of set A by set B is achieved by first dilating set A by B, then eroding the resulting set by B.

(iii) Thresholding function: change detection techniques act indiscriminately on the noisy portions and regions of interest of the image. In particular, when the *confiance* value does not allow decide if a region of

Figure 3: Example of *confiance image* after applying a median filter of size, respectively (from left to right): 3×3, 7×7, and 11×11.

Figure 4: Example of *confiance image* after applying the morphological closing operation for one iteration, three iterations, and five iterations.

interest is part of the figure or of the background, such techniques do not provide discrimination. The adopted approach to this problem consists in classifying the regions uncertain according to their proximity with pixels in which the value of *confiance* is 100% [32]. A threshold function is designed by means of a nonlinear convolution between the *confiance* image and a mask (*kernel*). The size and shape of the mask identify a neighborhood of image pixels. Hence, for K close to the pixel of coordinates (x, y),

$$\mathrm{dst}(x, y) = \begin{cases} 1 & \text{if } \mathrm{src}(x, y) \neq 0 \wedge \exists (u, v) : \mathrm{src}(u, v) = 1,\ (u, v) \ni K_{x,y}, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathrm{dst}(x, y)$ represents a pixel of the source image and $\mathrm{src}(x, y)$ a pixel of the image to which was applied the transformation. An uncertain pixel (with a value $\neq 0$) is not part of the background if in a neighborhood K it has at least a pixel with *confiance* value of 100%.

Figure 4 shows an example of *confiance image* after applying the morphological closing operation.

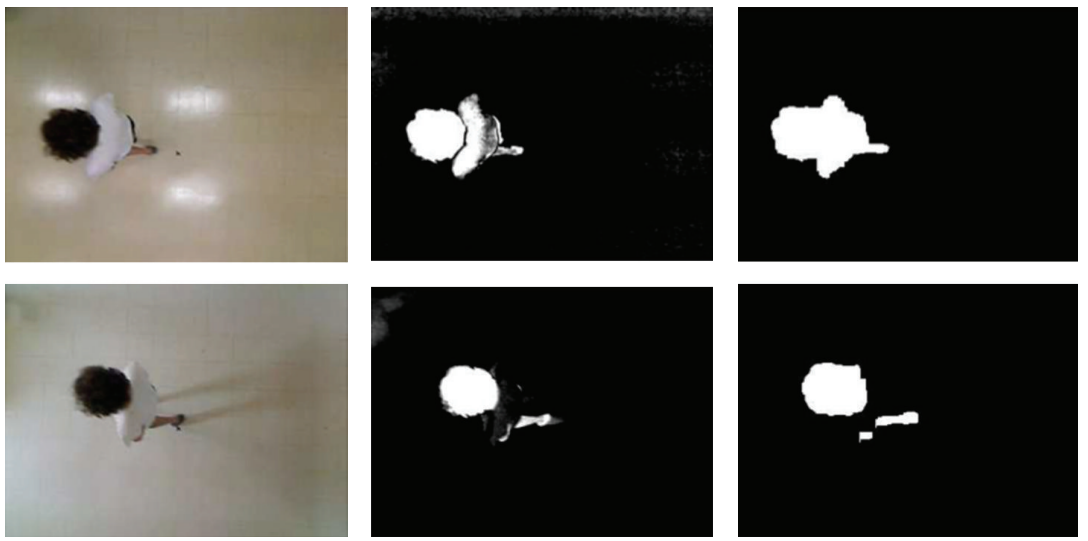
Figure 5 shows a frame, the *confiance image,* and the final image (blob) acquired in two real scenarios, without carpet on the floor and with/without artificial light: it can be noticed the robustness of the system, that permits individuating and tracking the person (the teacher, in this case).

Figure 6 shows the set-up panel of the platform with the sliders/buttons to control the software environment.

4.2. Applications. All the applications working in our platform do not require wearing sensors, handling pointing devices, or any tag to mark the environment: in this way the pupils (often with sensorial, neuropsychological, and relational impairments) are free to move their bodies without any constriction. The different applications can be used by the teachers depending on their pupils' needs. The open sound control (OSC) protocol enables the communication among the applications.

4.2.1. Two Applications for Blind People. The first one aims at offering blind users with an auditory *first sight* of space, an acoustic map of space, allowing them to immediately be aware of the type of environment and the objects it contains. The user—that does not need to perform a detailed tactile activity or a motor exploration and does not have to learn how to manage any kind of device—enters a room and, standing in the doorway, points in a direction in space with their arm. The system provides vocal information about the objects placed in that direction. These simple tasks aid blind people (1) to immediately recognize the type of environment, (2) to quickly detect a specific object, or (3) to safely move in space. This application allows the user to experience the environment both from an exocentric overview (e.g., a map) and an egocentric (user-based perspective) "view" [33]. The application uses a configuration file containing the information about the furniture of the room (the angular resolution is 2.3° in horizontal axis and 5° in vertical axis). A user-friendly interface allows the user to insert new objects. Because the low-cost hardware is used and the software is distributed for free, the application is used by (i) different Italian medical centers to facilitate the reintegration into private living spaces of a subject who acquired visual impairment in adolescence/adulthood and (ii) primary schools to facilitate the integration into the classrooms of blind pupils.

The second application (Figure 7) is an interactive multimodal system that extends the techniques of augmented reality (AR) to the domain of 3D audio and finds application



(a) (b) (c)

Figure 5: (a) A frame acquired by the camera; (b) the *confiance image* calculated by the system; (c) the final image (*blob*). The images are relative to an installation of our platform in a classroom, without carpet and with artificial light (above) and sunlight (below).

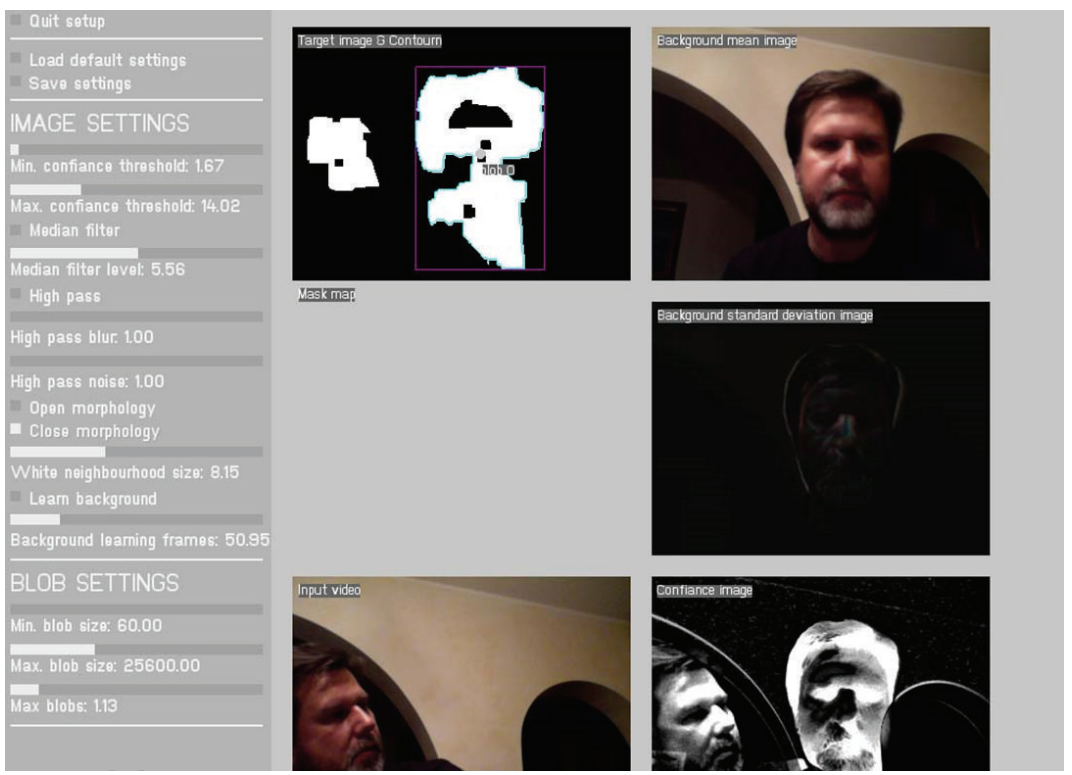


Figure 6: Set-up panel of the platform with the sliders/buttons controlling the software environment on the basis of the light conditions of the room.

exactly in this type of context positioning itself as a teaching aid for the blind, for the development and exploitation of the vicarious senses and the reduction of the secondary effects of blindness. With this application, the classical interaction modality, exclusively based on sight, has been replaced by hearing (with the reproduction of spatialized sounds) and touching (making use of tactile interfaces).

Pupils with visual impairments interact with the system by manipulating the cards. Here are the actions performed by the pupil during the game.

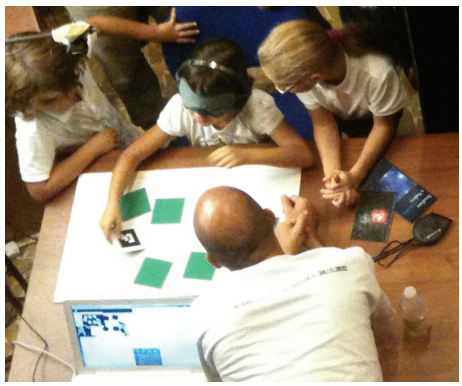


Figure 7: The application during the European Researchers' Night 2012, where it is tested by more than 2,000 users.

(1) Turn over the card and listen to the corresponding sound.

(2) Turn over another card and listen to its corresponding sound; the action is repeated until the paired sound is individuated.

(3) Confirmation from the system: the pupil presents both cards to the camera by drawing them together. If the choice is correct, the system plays a sound to confirm the success. If the choice is not correct, the system does not reproduce any sound: according to the experts in education of the blind and visually impaired, an error sound is a redundant information that could invalidate the task of memorizing the principal sound stimuli.

The application uses the optical tracking technology described in the last section and real-time audio spatialization techniques by means of four speakers in order to develop an acoustic-tactile interface, providing a play/educational dimension, accessible to blind and partially sighted users. The orientation of the card is not critical: the application is able to recognize the card with different orientations, to allow blind pupils to rotate the card.

Implementing these technologies allows users (i) to interface the system by means of real objects (cards with markers) and (ii) full freedom of movement (no need to wear sensors) within the three-dimensional sound space. The system has been tested on a heterogeneous sample of users by means of an usability test (following the methodology suggested in [34]) with satisfactory results in learning to read Braille, and the application is used by several institutes for blind people.

4.2.2. Two Applications for Speech Production/Discriminations. The first application provides users graphical visualization according to their speech production. It stimulates the user to use the voice features and movement to draw visual signs on a screen. The focus of this application is to exploit the most relevant features of speech and map them into graphic features. It is used as an assistive technology in the field of speech therapy in order to reinforce vocalization and speech skills (in primary schoolchildren with hearing impairments).

Figure 8: The graphical interface of the application aims to teach primary schoolchildren with hearing impairments the main features of voice: pitch, intensity, timbre, and duration through modifying the position, size, color, and the smile duration of the sun.

As illustrated in Figure 8, the application creates a graphical landscape on a screen; by clicking the graphical icons at the lower left side of the screen, pupils can change the landscape. The icon containing the microphone, at the lower right, is useful to calculate the noise environmental threshold, a very important function to discriminate silence from the users vocal sounds and to obtain the most efficient performance of the system possible, because the real school environment is usually rather noisy. The set-up icon at the lower right of the screen is useful to capture the users reference tone on which the application bases the extraction of the pitch. The application stores the value of the pitch in order to correctly visualize the graphical information. A yellow sun appears above the horizon when users make a vocal sound. By selecting the buttons on the right, users can choose which voice parameter visualize: the amplitude of sound is graphically represented by the variation in the size of sun which becomes bigger or smaller according to the magnitude of the sound pressure; the pitch corresponds to the variation in the position of the sun higher or lower above the horizon; the duration of sound is represented by a broad smile of the sun; the spectral content of the vocal sound is graphically visualized by a color variation (Figure 8).

The application have an user-friendly interface designed to keep the young users concentrated only to the system, avoiding distraction during the working session. The mapping between the auditory feature of vowel timbre and the visual feature of color is based on [35, 36] a study on biases in association of letters with colors across individuals both with and without grapheme-color synesthesia: even if, generally, the association of letters and colors is partially affected by environmental biases, the identified mapping of vowel and color is the following: red for /a/, green for /e/, blue for /i/, orange for /o/, and grey for /u/.

The ability to correctly and accurately interpret utterances produced by another person is essential for interpreting social interactions. Besides the difficulties with the possible social interactions, children with auditory processing concerns are likely to have severe difficulties in the class environment. On the basis of this assumption, an application for speech discrimination has been developed, using, as interface of the platform described in Section 4.1, a mobile device (iOS or Android tablet). It has been designed to measure, for each

Figure 9: The application for speech discrimination during the Researchers' Night 2012, where it is tested by more than 2,000 users.

Figure 10: The application for learning Italian as second language. The use of multimodality and the motor experience in learning increases the possibilities to build an effective and long-lasting knowledge.

sound feature, the child's ability (i) to discriminate the difference among a number of sound stimuli (sinusoidal as well as complex sounds), synthesized with different parameters (timbre, pitch, and intensity level), and (ii) to repeat the sound listened. The main goal of this application is to widen the scientific research about the auditory processing of sound in specific populations of pupils (such as children with Down syndrome) who, due to auditory processing disorders, encounter difficulties in speech production [36]. Moreover, this system allows the study, assessment, and intervention in pathologies and disorders of various etiologies that determine an impairment in speech production and articulation. This application is currently under testing in several preschools, in order to evaluate the speech discrimination skills of pupils. In addition, it was presented at the Researchers' Night 2012, where it is tested by more than 2,000 users (Figure 9). The European Researchers' Night is a mega event taking place every year on a single September night in about 300 cities all over Europe (http://ec.europa.eu/research/researchersnight/index_en.htm) in which each University presents its best research results.

4.2.3. Application for Learning Italian as Second Language. In order to help immigrants' sons to learn Italian language, an application was developed: the space captured by a web camera is divided into several areas, and sound or visual information corresponds to each area. The trajectory of the pupil's barycenter is used to match a sound to his/her specific position in space. A pupil explores the resonant space in which he/she can freely move without using sensors. Noises, sounds, and music are associated with peripheral zones and are reproduced when the child reaches and occupies a peripheral zone; audio reproduction of a story is associated with the central zone; the story provides references to the various sounds located in the peripheral areas. The child, listening to the story, enjoys searching for the sounds heard before, and, at the same time, he/she creates the soundtrack of the story. The use of this application offers pupils the possibility to enhance their communication skills providing alternatives or additions to the mode of communication already in place; encourage interaction with others and with the environment; extend their attention time. Figure 10 shows the application used by a pupil in a collaborative environment. Thanks to the total and immersive sound perception, the children have a general better understanding of Italian as well as an improved pronunciation and oral production. Even after some time, children are able to recall the exact contents learnt during a particular session within the application. The application is used in 12 Italian schools: from the video recording analysis, carried out with the teachers, we observed that this method of teaching increases the motivation to listen and consequently to learn new words and phrases. The evaluation tests carried out successively by the teachers show that pupils master the contents assimilated through the application.

5. Discussion and Conclusions

The real challenge for educators today is to organize learning environments and teaching practices that ensure effective learning. Given that learning is a very complex process, different cognitive perspectives have to be considered [35]. Significant increases in learning can be accomplished through the use of visual and verbal multimodal systems: "students engaged in learning that incorporates multimodal designs, on average, outperform students who learn using the traditional approaches with single modes" [37].

The use of multiple representations of knowledge, particularly in computer-based learning environments, has now long been recognized as a very powerful way to facilitate learning [1]. Multiple representations usually involve the use of PowerPoint presentations as minilectures, text synchronized with images, interactive diagrams, video presentations, audio explanations of concepts, and images. In these learning environments, the multimodal elements (visual, auditory, and interactive) are merely presented as additional representations of information, and thus the level of interactivity is very low.

This work illustrates a platform (see Section 4.1) that groups some applications (described in Section 4.2), aimed at implementing specific learning methodologies particularly focused on pupils with impairments. These applications are focused on specific educational objectives: assisting people with visual impairments in orientation, mobility, and navigation skills acquisition; proving information on the acoustic properties of voice; enhancing assessment techniques of speech discrimination. We are distributing for free the system (platform and applications) to tens of Italian primary schools. All the teachers involved in this experimentation phase, sometimes together with the pupils themselves, developed

the multimedia content (often inspiring to fairy tales) to enrich and customize the applications, by means of a (very) user-friendly interface of the platform, exactly designed to be used by people without informatics skills.

Each of the above applications was evaluated by means of various usability testing focused on measuring the capacity of the systems to meet the intended purpose. The results of the different experimentations are highlighting the weak points, allowing to find solutions and to improve the overall performance of the system. For the results of the first experimental phase, see [38, 39].

The continuative use and the experimentation of these applications in real educational context provide effective engagement in learning, when

(i) tasks are carefully planned in a logical, coherent sequence;

(ii) the interactive activities are integrated with curriculum content and skills;

(iii) systems are able to answer to and use different communication channels.

The educational possibilities that this type of technological environments might be implemented, if integrated with curriculum contents and skills, are endless; to date, there are few similar experiences, and therefore, research on interactive and multimodal environments for learning in real educational contexts is a field that has to be explored. However, the positive results of the research documented in this paper suggest that, even if much work has still to be done, the premises are encouraging. The results of the scientific experimentations of the above specific multimodal and interactive systems demonstrated indeed that the use of multimodality and the motor experience in learning increases the possibilities to build an effective and long-lasting knowledge.

Following our approach, we are developing other applications, in particular in the field of specific learning difficulties and special educational needs. Indeed, the educational context has to face with an increasing number of students with specific learning difficulties and disabilities which typically affect the student's motor skills, information processing, and memory. Teachers, in order to deal with this complex situation, require to implement "inclusive teaching" which means recognizing, accommodating, and meeting the learning needs of all the students. However, teachers' experience has demonstrated that adjustments made for students with disabilities can very often benefit all students.

Summarizing, the specific objectives of our research are as follows.

(i) In the scientific field:

(a) contribute, by providing data, to research in the field of multimodal interactive environments for learning;

(b) develop enactive interfaces which are able to provide a multimodal input and output (plurisensorial);

(c) contribute to the comprehension of the relationship between the emotional answers of the user and the sensorial experiences which occur inside the multimodal environment;

(d) develop techniques for the automatic extraction of high-level features from a multimodal input;

(e) set a participant design approach in the development of the environment in which the users act as codesigners and where the design process entails also a learning process in which the designers and the users learn together;

(f) develop a system of assisted technology to use in the field of (i) training of the listening comprehension in various populations of pupils, both Italian and immigrants, having different disabilities and (ii) assessment of speech production and discrimination. The research project is focused on pupils with sensorial, neuropsychological, and relational impairments;

(g) design of multimodal interactive systems to improve the relational skills which due to social and family difficulties or to the presence of particular syndromes—that is, Down syndrome—are hindered.

(ii) Social and educational field:

(a) proposing a multimodal interactive systems for learning and communication aims to

(1) supply an alternative and/or additional tool to promote different cognitive styles;

(2) suggest compensatory and dispensatory measures in cases where, because of situations of disability, the traditional methods of teaching do not allow the users to follow an appropriate learning pathway or to completely take part in the educational environment;

(3) create a sensitive environment with a strong interactive component which involves the movement of the entire human body, the expressiveness of gestures, and the use of all five senses;

(4) propose an alternative technology for second language learning (e.g., Italian as second language, for immigrant);

(5) reconsider the motor aspect of knowledge and implementing it to face learning difficulty situations;

(b) promote learning motivation, improve communication and relational abilities by encouraging the interaction, extend the attention span, and improve personal autonomy.

A future development will be the creation of "dynamic environments" which allow the user to interact with them altering their structure and features. The goal is to implement user-world and multiuser interactions which allow the user to

interact with the real world in a much more engaging way. In this kind of environments, learners can organize resources, manipulate information, and even create new content also in collaboration with others. Within dynamic environments, students are not simply consumers of information, but they become part of an active learning experience. Another very interesting direction that we are following is the development of techniques for the measurement of nonverbal social signals. The exploitation of nonverbal social interaction in interactive and multimodal systems can be very useful to support effective learning and cocreation.

Acknowledgments

The authors would like to thank the research group coordinated by Chiara Levorato (Department of Developmental and Socialization Psychology, University of Padova, Italy) for the psychological guidelines helpful for the definition of system requirements; Serena Zanolla (teacher in "E. Frinta" Primary School, Gorizia, Italy) for her help in validation processes; Leonardo Amico, Giorgio Tempesta, and Daniele Marabese for the programmer work. This work was partially supported by the projects (i) *Evolution of Acoustic Feedback Systems for the Homecoming of Subjects with Acquired Vision Impairment*, funded by the Friuli Venezia Giulia Region (Italy), Decree n. 1265/AREF dd. November 25th, 2010 and (ii) *MILE: Multimodal and Interactive Learning Environment*, funded by Veneto Region (Italy), FSE 2007/2013—Human Capital Axis, Dgr n. 1686.

References

[1] R. Moreno and R. Mayer, "Interactive multimodal learning environments: special issue on interactive learning environments: contemporary issues and trends," *Educational Psychology Review*, vol. 19, no. 3, pp. 309–326, 2007.

[2] D. H. Jonassen, "Computers as mindtools for engaging critical thinking and representing knowledge," in *Proceedings of the Educational Technology Conference and Exhibition (EdTech '99)*, 1999.

[3] L. Vygotsky, *Thought and Language*, MIT Press, Boston, Mass, USA, 1986.

[4] S. Price and Y. Rogers, "Let's get physical: rhe learning benefits of interacting in digitally augmented physical spaces," *Computers and Education*, vol. 43, no. 1-2, pp. 137–151, 2004.

[5] J. Bruner, *Toward a Theory of Instruction*, Belknap Press of Harvard University Press, 1966.

[6] A. Karmiloff-Smith, "Nativism versus neuroconstructivism: rethinking the study of developmental disorders," *Developmental Psychology*, vol. 45, no. 1, pp. 56–63, 2009.

[7] S. Bhupinder, K. Rupinder, D. Nidhi, and K. Ramandeep, "The process of feature extraction in automatic speech recognition system for computer machine interaction with humans: a review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 2, 7 pages, 2012.

[8] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, 1999.

[9] M. Puckette and T. Apel, "Real-time audio analysis tools for pd and MSP," in *Proceedings of the International Computer Music Conference*, pp. 109–112, San Francisco, Calif, USA, 1998.

[10] C. Sinthanayothin, N. Wongwaen, and W. Bholsithi, "Skeleton tracking using kinect sensor & displaying in 3D virtual scene," *International Journal of Advancements in Computing Technology*, vol. 4, no. 11, pp. 213–223, 2012.

[11] K. Akita, "Image sequence analysis of real world human motion," *Pattern Recognition*, vol. 17, no. 1, pp. 73–83, 1984.

[12] J. O'Rourke and N. I. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 522–536, 1980.

[13] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8–15, Santa Barbara, Calif, USA, June 1998.

[14] T. Horprasert, I. Haritaoglu, C. Wren, D. Hardwood, L. Davis, and A. Pentland, "Real-time 3D motion capture," in *Proceedings of the 2nd Workshop on Perceptual User Interfaces*, 1998.

[15] H. D. Yang and S. W. Lee, "Reconstruction of 3D human body pose from stereo image sequences based on top-down learning," *Pattern Recognition*, vol. 40, no. 11, pp. 3120–3131, 2007.

[16] J. Rodgers, D. Anguelov, H. C. Pang, and D. Koller, "Object pose detection in range scan data," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2445–2452, June 2006.

[17] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 755–762, San Francisco, Calif, USA, June 2010.

[18] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '10)*, pp. 3108–3113, Anchorage, Alaska, USA, May 2010.

[19] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vision Computing*, vol. 1, no. 1, pp. 5–20, 1983.

[20] L. Snidaro, G. L. Foresti, and L. Chittaro, "Tracking human motion from monocular sequences," *International Journal of Image and Graphics*, vol. 8, no. 3, pp. 455–471, 2008.

[21] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "P finder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[22] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[23] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[24] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proceedings of the NIPS*, pp. 1097–1104, 2004.

[25] L. Xia, C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '11)*, pp. 15–22, Colorado Springs, Colo, USA, June 2011.

[26] H. M. Zhu and C. M. Pun, "Movement tracking in real-time hand gesture recognition," in *Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science (ICIS '10)*, pp. 240–245, Yamagata, Japan, August 2010.

[27] M. R. Malgireddy, J. J. Corso, S. Setlur, V. Govindaraju, and D. Mandalapu, "A framework for hand gesture recognition and spotting using sub-gesture modeling," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 3780–3783, Istanbul, Turkey, August 2010.

[28] Y. Gu and C. Yuan, "Human action recognition for home sensor network," in *Proceedings of the International Conference on Information Technology and Software Engineering*, vol. 212 of *Lecture Notes in Electrical Engineering*, pp. 643–656, 2013.

[29] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, 2008.

[30] T. Lei, Y. Wang, Y. Fan, and J. Zhao, "Vector morphological operators in HSV color space," *Science China Information Sciences*, vol. 56, no. 1, pp. 1–12, 2013.

[31] N. Falco, M. D. Mura, F. Bovolo, J. A. Benediktsson, and L. Bruzzone, "Change detection in VHR images based on morphological attribute profiles," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 636–640, 2013.

[32] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and location of people in video images using adaptive fusion of color and edge information," in *Proceedings of the 15th IEEE International Conference on Pattern Recognition*, vol. 4, pp. 627–630, 2000.

[33] R. Ramanathan, *Combining egocentric and exocentric views: enhancing the virtual worlds interface [Master's Project]*, Department of Urban and Regional Planning, University of Illinois at Urbana-Champaign, 1999.

[34] M. Obrist, F. Förster, D. Wurhofer, M. Tscheligi, and J. Hofstätter, "Evaluating first experiences with an educational computer game: a multi-method approach," *Interaction Design and Architecture*, vol. 11-12, pp. 26–36, 2011.

[35] M. Steward, "Learning through research: an introduction to the main theories of learning," *JMU Learning and Teaching Press*, vol. 4, no. 1, pp. 6–14, 2004.

[36] M. Roch and M. C. Levorato, "Simple view of reading in Down's syndrome: the role of listening comprehension and reading skills," *International Journal of Language and Communication Disorders*, vol. 44, no. 2, pp. 206–223, 2009.

[37] C. Fadel and C. Lemke, "Multimodal learning through media: what the research says," Tech. Rep., Cisco Systems, 2008.

[38] S. Zanolla, S. Canazza, A. Rodà, and G. L. Foresti, "Learning by means of an interactive multimodal environment," in *Proceedings of the SETECEC Conference*, Alaipo and Ainci, Eds., pp. 167–176, Venice, Italy, March 2012.

[39] S. Zanolla, S. Canazza, A. Rodà, A. Camurri, and G. Volpe, "Entertaining listening by means of the Stanza Logo-Motoria: an interactive multimodal environment," *Entertainment Computing*, 2013.