Natural Interfaces to Enhance Visitors' Experiences

Thomas M. Alisi, Alberto Del Bimbo, and Alessandro Valli
University of Florence

Museums and exhibitions don't communicate. These places are often just a collection of objects, standing deaf in front of visitors. In many cases, objects are accompanied by textual descriptions, usually too short or long to be useful for the visitor. In the last decade, progress in multimedia has allowed for new, experimental forms of communication (using computer technologies) in public spaces. Implementations have ranged from simply using standard PCs with multimedia applications that show thumbnails of image data integrated with text to large theaters that immerse users in virtual worlds or reproduce and display 3D models of masterpieces. Often designers just apply the technology available to traditional museum schemes, without paying much attention to the visitor's experience, particularly to the ways they expect users to interact with the system or to the cognitive and aesthetic factors involved.

Natural interaction, on the other hand, can reduce the gap between computing and ordinary physical things. But it requires that interfaces differ from traditional human–computer graphic interfaces based on menus, icons, a mouse, and a keyboard. They also must differ from advanced user interfaces such as those used in augmented and virtual reality applications, including special wearable devices such as data helmets, glasses, and gloves; body markers; and smart card technologies, which enable higher user mobility.

In this article, we discuss two multimedia system installations that we designed to provide information to museum and exhibit visitors through natural-interaction interfaces. The Media Integration and Communication Center at the University of Florence used natural-interaction, computer vision systems for both

- the Point At system in the Museum of Palazzo Medici Riccardi in Florence (a permanent installation that communicates the content of art masterpieces) and
- the Interactive Blackboard system at the New City Exhibition of Palazzo Vecchio in Florence (for a January 2004 display of the Florence municipality projects).

These systems let users communicate with unaugmented hand gestures, without the need for special tools or training.

Editor's Note

The authors present a multimedia system that really works in a cultural public space. Indeed, if you go to Florence and visit the museum of Palazzo Medici Riccardi, you might see a queue of worldwide tourists waiting for their turn to play with a digital version of the famous fresco *The Journey of the Magi*, appearing on two large screens. Visitors stand in front of the screens and point with their hands to the part of the painting they're interested in. Two cameras grab this point and an algorithm calculates the exact part of the painting the person selected. In response to the pointing, an audio response gives information on the subjects or objects.

Visitors seem to deeply enjoy their interaction with the system, which does feel *natural*. Visitors wear no special equipment and use no complex hardware; the fresco is extremely well displayed, and typically the information is precise and interesting, with different levels of information available.

—*Tiziana Catarci*

Project goals

We established the basic natural-interaction system framework for our two installations in Florence with four design principles in mind:

- We equipped computers with appropriate hardware and software that senses people's actions and intentions, at both perceptual and interpretative levels.
- We based the interaction schemes on natural interaction between people and between people and physical objects, as in everyday life.

1070-986X/05/$20.00 © 2005 IEEE

- The technology disappears into the environment, empowering ordinary objects and spaces—thus, removing the (frustrating) idea that users are dealing with a computer.

- Our interactive design pays close attention to cognitive aspects and aesthetics.

The projects' goal was to open the door to a new generation of artifacts that users can easily and intuitively exploit, shifting their attention from the interface to the content.

To do this, natural-interaction interfaces must respect the typical behaviors between humans and let the user interact with the system almost in the same way. This type of interaction is now possible thanks to advancements in computer vision and speech processing. In particular, computer vision makes it possible to interpret image content and implement robust, environment-independent tracking methodologies to develop effective human–computer interfaces. For our work, we chose vision-based hand-pointing systems because hand pointing is an everyday activity, so it doesn't require any a priori skills or training.

Point At

Palazzo Medici Riccardi is one of the most important museums in Florence. It was the former house of the Medici family, and in the Medicis' small private chapel, it hosts one of the masterpieces of Renaissance painting, the famous fresco *The Journey of the Magi* by Benozzo Gozzoli (1421–1497). To manage the large number of visitors and preserve the frescoes, the museum only admits seven visitors into the chapel at one time, for at most seven minutes. In this small time period, visitors often spend most of the time trying to find correspondences between the descriptions of the fresco contained in guidebooks and the painting content. Because the fresco contains so many characters, visitors often overlook the fresco's beauty, structure, and significance.

This motivated the museum to find a new method to teach the visitors about the fresco's content, before visiting the chapel. The ultimate goal was to use information technology to make teaching attractive and effective. The Point At system's goal is to stimulate the visitors to interact with a digital version of the fresco and, at the same time, make them interact in the same way they will in the chapel, reinforcing their real experience with the fresco.

Lorenzo il Magnifico's bedroom, located on the palace's ground floor, hosted the teaching environment. We installed two large screens that project two digital versions of the fresco. Visitors are invited to stand in front of the screens and indicate with their hand the part of the painting that interests them. Two digital cameras grab the visitors' pointing action and a computer vision algorithm calculates the screen location where they're pointing. The system then provides audio information about the subject or object.

The basic idea of this interactive environment is to let visitors replicate the gestures and behaviors that they would use in the chapel to ask a guide something about the fresco such as, Who is that character? or Why is he/she in the fresco? In designing the system, we considered the following issues:

- *Easy and simple interaction*. Visitors shouldn't need any instruction or have to wear any special device.

- *High-resolution display*. The fresco must be displayed so that visitors can appreciate even small particulars (almost invisible in the real chapel).

- *Culturally sound and meaningful explanations*. The text must explain the painting's social, political, and cultural context so as to give more complete contextual information. In addition, it must be short enough to be stimulating but not boring.

- *Interactivity for different categories of visitors*. Interaction should be satisfactory for visitors who just want an idea about the fresco, for those who are attracted by particular characters, or those who want to have complete information on the whole fresco.

- *Pleasing physical setting for different categories of visitors*. The physical setting must host both active and passive visitors (for example, the relatives of the person who's actually interacting with the system and those interested in listening but not in being active).

- *Good aesthetic presentation*. The interactive environment must be integrated within the museum and must respect the visitors' whole experience. Audio and video should be comparable to cinema and state-of-the-art

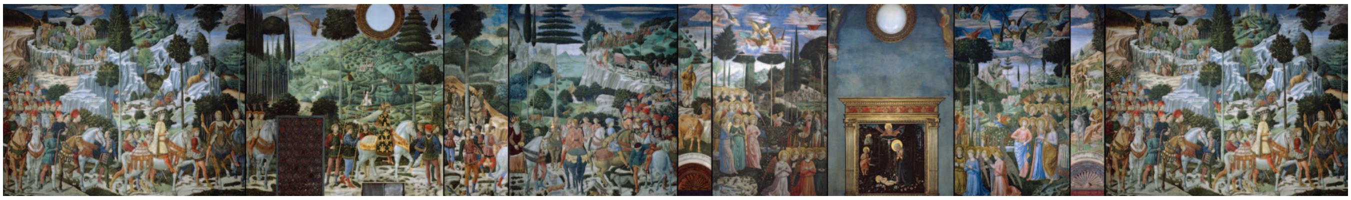

Figure 1. The whole stripe of Benozzo Gozzoli's fresco **The Journey of the Magi.**

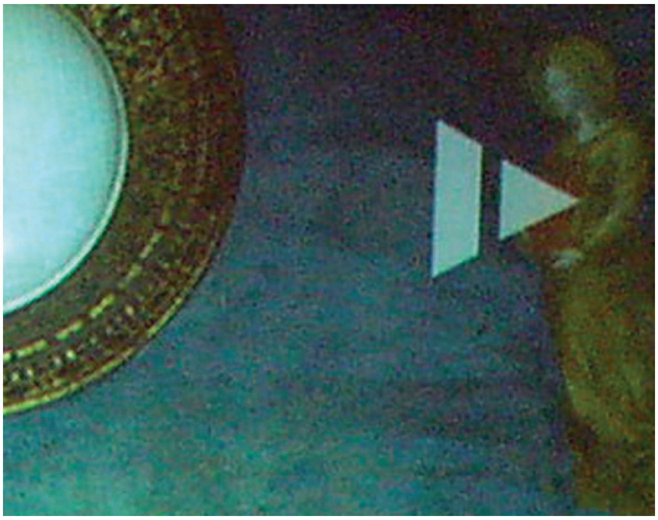

Figure 2. The scroll right button.

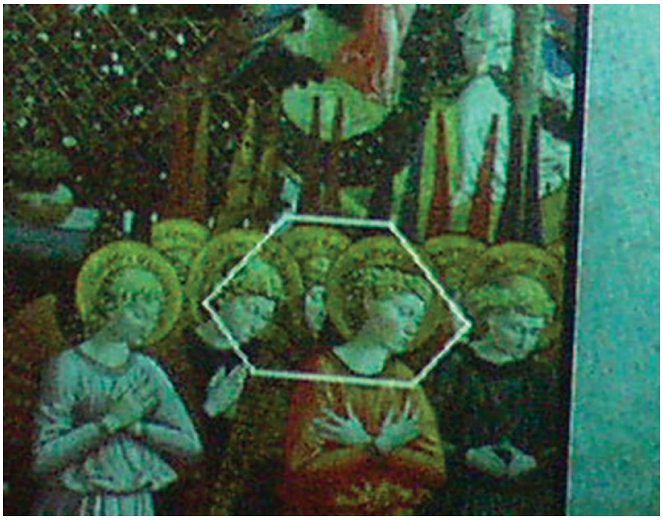

Figure 3. This graphic sign identifies important particulars associated with zooms and explanations.

computer games, although at the same time seriously support the visitors' understanding of the art content.

Finally, we designed the system to work unattended in a real-world setting.

Content definition

Professional photographers took numerous images of the whole fresco, which we then digitized at high resolution—the smallest particular fits a 1,024 × 768 screen. We edited them to obtain a continuous horizontal image stripe (see Figure 1). Art historians had a hard time carefully selecting which among the large number of characters and objects were most worthy to explain to a general public. They provided appropriate textual descriptions for each of the characters and objects they selected. Each text follows a precise pyramidal structure: a brief description, followed by one or two detailed explanations.

The texts, which we translated into different languages, were read and recorded by professional speakers. Because of the text's structure, visitors can listen to the beginning to get a general idea of the subject and then stop and proceed to other subjects, or they can listen to the end of a speech for a deeper understanding. We limited the audio length per character or object to a maximum of 45 seconds to favor fast-paced and intriguing material. A composer prepared music pieces inspired by Renaissance music for background sound.

Designing natural interactivity

Visitors stand in front of large, rear-projected displays (approximately 2.5 meters wide). The distance between the pointing person and the screen is approximately 2.5 meters. Only one person at a time can interact with the screen, so we installed two screens to permit two visitors simultaneous access. The screen displays only one portion of the fresco, but users can inspect the whole fresco by pointing at two side icons in the form of arrows (see Figure 2) and scrolling horizontally. The system highlights the 26 most important characters and objects with a thin graphic drawing so that the visitor can easily identify those figures that have some explanation associated with them (see Figure 3). Visitors can select each of them by simply pointing at it with their hand for at least one second. The system doesn't require any external device.

When a visitor points at a character or object, the system magnifies it and displays a text label that shows the character or object's name, thus precisely contextualizing the subject's information (see Figure 4). At the same time, the visitor hears the recorded audio for the character or object. (Visitors can select a language for the audio with physical buttons when they first approach the display.) The zoomed image includes a back button; by pointing at it, the visitor stops the audio, zooms out, and goes back to the original portion of the fresco. We designed the zoom and pan's acceleration and deceleration to give a seamless sensation, with no sudden changes, to keep the visitors' attention and, at the same time, let them understand the interface's behavior. No traditional GUI elements, like windows or scrollbars, are present. We reduced the necessary buttons as much as possible to avoid an invasive display. In this way, we preserve the fresco's beauty so the visitors' attention is fully devoted to the fresco content.

Because of space limitations, we don't discuss the scientific aspects of this system, including the geometrical solution of the pointing problem along with the processing algorithms here. (See related literature for more details.[1])

Visitor experience

We installed the Point At system at the Palazzo Medici Riccardi on 10 December 2003 (see Figure 5). Since that time, it hasn't malfunctioned or experienced service interruptions. Visitors' reactions have been generally enthusiastic; most people use the system intuitively or by imitation—after watching someone else interact with it.

From the system's log, we derived statistical observations on the number of users, the mean number of characters a single person explores, the mean number of seconds listened for each character, and so forth. Interestingly, visitor behavior is heterogeneous, so we can't make any general assumptions. For example, some people stand at the system for less than two minutes and others explore the whole fresco to hear about all the characters (about 10 minutes). The number of users with the most interest and the number of those who listen to the explanations for two to four minutes are almost equally distributed.

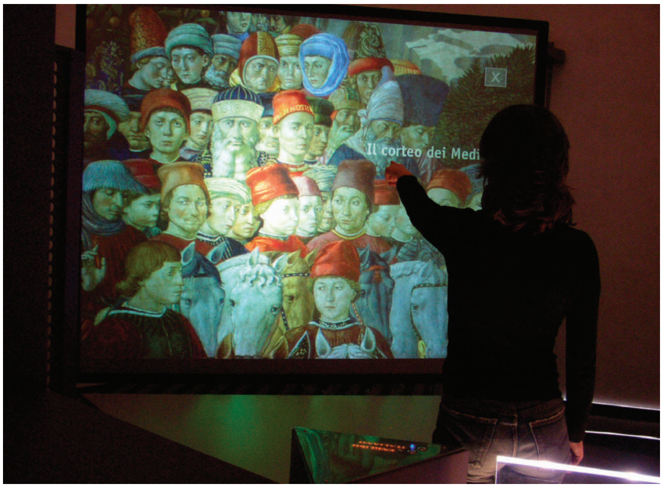

Figure 4. Learning about a group of characters in Gozzoli's fresco.

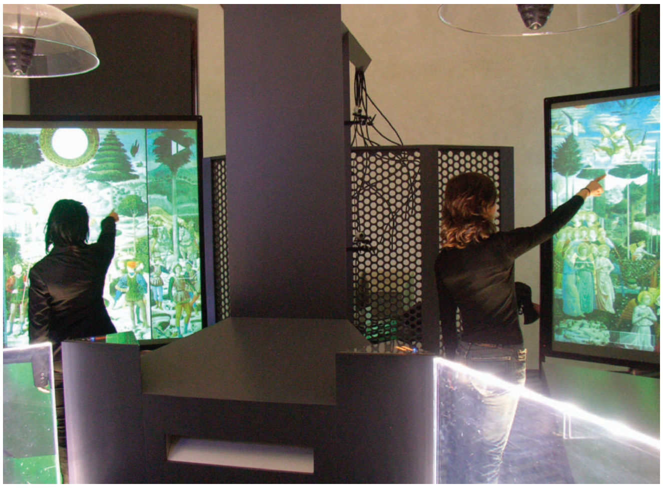

Figure 5. Two users at the Palazzo Medici Riccardi interacting with the Point At system.

Interactive Blackboard

In January 2004, the Florence city council promoted a two-month-long event to inform its citizens about the transformations that Florence was going through—particularly, new strategic infrastructures, the restoration of large areas, and new services. The event included temporary exhibitions and a permanent location where the most important initiatives were presented with movies, models, and interactive kiosks. The designers decided a system needed to show citizens the new urban transportation system (line paths and stations) and use some novel solutions to attract visitors to interact with the system, offering at the same time easy and immediate operation.

The resulting system, the Interactive Black-

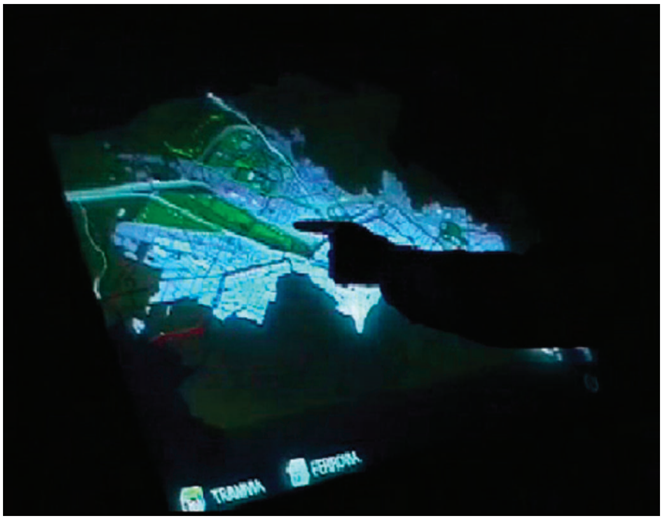

Figure 6. The Interactive Blackboard displaying Florence's urban area.

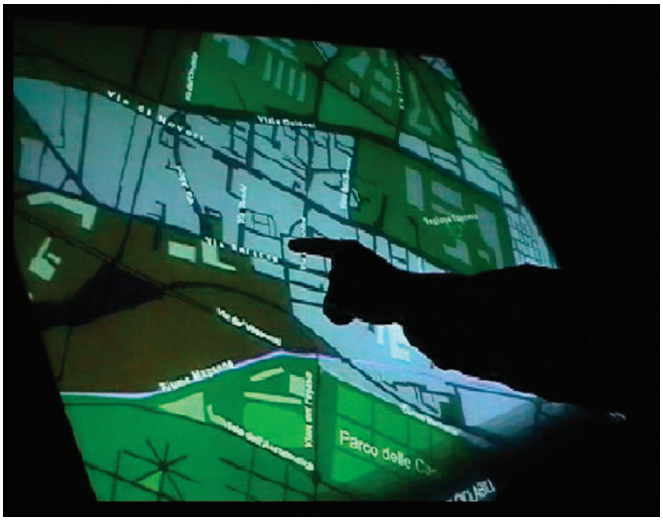

Figure 7. Example of a user choosing a specific street in Florence with the Interactive Blackboard.

board, consisted of a large rear-projection screen, 1.2 meters wide, which displayed a colorful 3D map of Florence, at four zooming levels. In particular, users could move from a global view of the whole urban area (see Figure 6) to a detailed view of a few streets in a certain city neighborhood (see Figure 7). The map had a clean design so users could find the area around their homes and find the closest train station or tramway stop. Buttons in the lower part of the screen let users zoom out and select the kinds of information to display—whether related to trains, buses, cycle routes, or parking spaces.

We designed the Interactive Blackboard's screen so that users only have to move their hand to point at a zone on the map to zoom in and out or to the icons they want to switch between the different information layers. The user's pointing action is grabbed by a single digital camera, looking down at the inclined screen, positioned over the blackboard, high above the user's head (about 2.5 meters above the floor level). The system computes the zone the user points to in real time, after image processing and interpretation. Similar to the Point At system, the Interactive Blackboard triggers when the user points for about one second, with no visual feedback on screen. Although the user doesn't need to touch the screen, the effect isn't different from that of a giant touch-screen display. Because the exhibition space was crowded and noisy, we only used the audio channel to play simple and neat sounds to confirm the user's actions.

Technical implementation

The Interactive Blackboard takes images in the near-infrared spectrum—the system had to operate in a dark indoor space—at a resolution of 320 × 240 pixels processed at 30 frames per second. We placed two near-infrared illuminators on top of the screen. The light rays were parallel to the screen surface to illuminate the users' hands near the screen surface without lighting the screen. In this way, we could easily separate the pointing hand from the background, estimate the fingertip position on the screen plane, and transform it into display pixel coordinates after a 2D-to-2D perspective model transformation.

We segmented the image into background and foreground regions according to statistical matching with an adaptive model of the observed scene and removed noise using common topological filters. Foreground pixels (corresponding to the user's hand, arm, and body parts) were then clustered into blobs, using a connected components algorithm. For each blob, we computed simple descriptors, such as moments (up to third order) and extreme points.

We needed perspective calibration (plane to plane) to compute the interface screen coordinates from the camera image plane coordinates. Figure 8 shows the geometrical model we used to solve the problem of a user pointing to a tilted screen with the hand close to the surface.

The relation between the pointing fingertip (u,v) on camera's image plane $\prod_i$ is related to the corresponding screen coordinates (X,Y) in pixels on $\prod_s$ by a planar homograph transformation that using the projective camera model has the explicit form

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \qquad (1)$$

with eight independent parameters.

Writing the H homogeneous matrix in vector form h, the homogeneous equations for each point become

$$\begin{bmatrix} u & v & 1 & 0 & 0 & 0 & -uX & -vX & -X \\ 0 & 0 & 0 & u & v & 1 & -uY & -vY & -Y \end{bmatrix} h = 0 \qquad (2)$$

We can estimate h directly from the matrix's singular value decomposition obtained by the constraints in Equation 1 for at least four points.

The segmentation algorithm was the system's most critical element. Its effectiveness depends on the right choice of the threshold used to discriminate between foreground and background pixels. We implemented the Otsu algorithm to compute the histogram of the intensity distribution over an image region and estimate the statistically best threshold that separates the distribution in two coherent spaces. According to this, a single threshold doesn't exist for the whole image, but the thresholding algorithm uses a dynamic threshold that changes in time and space.

We computed the fingertip position by considering the extreme points of the contour of the foreground blobs, exploiting the information obtained from their tracking, and considering some heuristics that we derived from the system layout—for example, users always approach the screen surface from a given direction.

Future work

The two prototypes we present here open the door to a new family of interactive devices seamlessly integrated into an environment. Such devices will flourish in cultural exhibitions and retail showrooms, allowing easy and satisfactory interaction with digital media content.

We're currently working on letting multiple users access the system simultaneously. We're also focusing on real-time analysis of visitors' behaviors in front of interactive artifacts. This work will move us from a simple reactive system to fully proactive system. It will soon be possible to have systems greet passersby, encourage shy visitors, and explain to passive users how to interact. **MM**

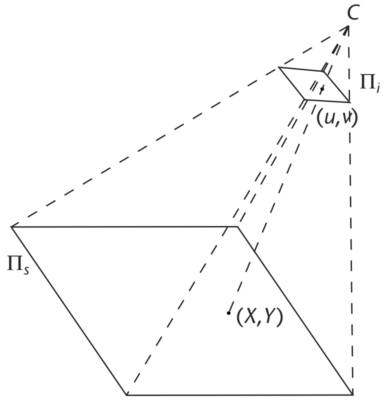

Figure 8. The geometrical model used in the Interactive Blackboard.

Reference

1. C. Colombo, A. Del Bimbo, and A. Valli, "Visual Capture and Understanding of Hand Pointing Actions in a 3D Environment," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 33, no. 4, 2003, pp. 677–686.

Readers may contact Albert del Bimbo at delbimbo@dsi.unifi.it.

Contact department editor Tiziana Catarci at catarci@dis.uniroma1.it.