A Prototype System for Intuitive Film Planning

Michael Hoch
Academy of Media Arts
Peter-Welter-Platz 2, 50676 Cologne, Germany,
e-mail: micha@khm.de
www.khm.de/~micha/Projects/research.html

Abstract

In this paper we describe a prototype system for directing a computer generated scene for film planning. The system is based upon the concepts of the Intuitive Interface, an environment in which the user interacts in front of a projection screen, and where interaction in physical space and pointing gestures are used to direct the scene. The environment demands certain requirements of the interaction paradigms that cannot be converted directly from menu-based systems. We present some new concepts using real space that cannot be found in traditional desktop metaphors. For the pointing gesture we describe a simple algorithm based on tracking colored regions and inverse kinematics using a simple model of the human arm. Finally, we will discuss some results from an empirical study with film directors using the prototype.

1 Introduction

The Intuitive Interface is a research project that investigates new methods of interaction with the computer by freeing the computer from the non-flexible desktop setup and integrating the interface in the every-day physical environment of the user [9]. We focus on creative people, as users, who are often unfamiliar with computers and hesitate before getting involved with mouse and keyboard, because of the symbolic and alphanumeric abstraction that is necessary. The system makes strong use of body movements to trigger commands and real space in front of a rear projection screen to store and retrieve information (Figure 1). This so called memory-function is based on a technique called "Ars Memoriae" which was used by orators in antiquity to memorize long speeches [16]. The orator would create icons of the subjects to be memorized and place them at chosen locations in an imaginary architecture. The icons are used as links to the information and as the orator (in his mind) walks to the memorized locations he is able to retrieve the icons and with them the information. Our scenario uses this metaphor by allowing the user to put entities at chosen locations in real space and to retrieve these icons and the linked information at a later time.

A stereo computer vision system is used for sensing user posture and simple gestures like pointing, together with a speech recognition software that is used to trigger commands. The vision system is based on color segmentation and blob analysis.

Figure 1: Intuitive Interface with film planning application

In film production storyboards are normally used to plan a film. Storyboards are drawn sketches of the key scenes of a sequence. These sketches show camera perspective and give a good description of the scene setting, but they lack the final impression of the image, and movements cannot be shown directly. In our application, the system is used for directing a computer generated scene for film planning purposes [6]. The example scene shows a typical set of a room (Figure 1). Objects or figures in the scene can be moved around by simply pointing to them and issuing a speech command. By moving in real space, the user can call up a stage setup for adding objects to the scene. Furthermore, modifying camera position and point of view is performed by changing the position in real space. Camera movements can also be defined, recorded, and played back by moving in front of the scene. The entire space in front of the display is used to direct the scene. This is a familiar working environment for film people as opposed to the desktop environment of traditional PC. Instead of a storyboard, the result is a sample sequence of the film. For a more detailed description of the concepts see [6].

In the following we will first present related work. Next we discuss the interaction requirements of our environment and compare them to traditional WIMP interfaces. There-

after, we will describe the prototype system. Finally, we will comment on our results and some problems.

2 Related Work

Using body movements to control game-style applications is described in [1]. We extend this approach to an application in the creative field, such as an interactive film planning system [6] in which the real space is used as a location where data is placed and retrieved, and not only as a relative reference system for navigating in a virtual space. An "intelligent room" is described in [13] that supports the user's daily activities in an office or meeting environment. In our approach, space and gestures are not only used to trigger commands, but real space becomes part of the computer's data space and the body movements in this space become part of the interface. Using two-dimensional graphic space for data-management is described in [2], using real space as a location to situate windows of a standard window system is described in [5]. We extend these approaches by freeing the user from tracking devices and allowing her to use the entire room to store any kind of information.

Unlike other gesture recognition systems, we do not seek to recognize complex gestures which, although they often allow great functionality, have a large learning curve and are sometimes awkward to use (for example the different hand gestures in [3, 11]). Our commands are as simple as pointing gestures and more attention is given to the users body movements in real space.

An approach for tracking the human body has been presented in [15]. The limitations of the Pfinder approach are the processing speed that is about 10 frames per second for a single image and the lack of multiple user support. For our application, we found that we need at least 12 fps for a stereo system, to give enough feedback to the user while performing the pointing task. For this reason, we have chosen a much simpler approach for tracking the user based on colored regions. We also support multiple users in our environment.

3 Interaction Requirements

The Intuitive Interface places the user in an unencumbered environment in front of a rear projection screen. This special setup induces special requirements on the interaction. Interaction paradigms that have been proven to be useful in menu-based systems, might not work in our environment, i.e., they cannot be converted directly. We found that most of the differences stem from one of the following:

- pointing to a large screen is more a deictic gesture and not as exact as using the mouse
- the large projection screen has a different look and feel from a computer monitor, resulting in a different impression when, for example, menus are shown
- real space and body movements can be used as part of the interface. They have, on the other hand, no counter part in desktop systems

These findings result in different demands of object and menu selection techniques and new concepts using real space that cannot be found in traditional desktop metaphors.

3.1 Pointing and Selecting

Pointing to objects shown on a projection screen is a different task from traditional mouse pointing. It is more related to pointing-like gestures in social communication. In face-to-face conversation, for example, humans frequently use deictic gestures (e.g., the index finger points at something) parallel to verbal descriptions for referential identification. Unlike the usual semantics of mouse clicks in direct manipulation environments, in human conversation the region at which the user points is not necessarily identical with the region to which he or she intends to refer [12]. Natural pointing behavior is often ambiguous or vague. Therefore, a desktop application cannot be directly converted to an application in our environment.

For testing the pointing gesture we displayed a Netscape window on the projection screen [8]. Mouse movements where simulated by analyzing the pointing direction of the user and setting the mouse to that position. Mouse clicks could be issued by a speech command. The screen dimensions were 2 by 2.6 meters, and the user acted from a distance of between 3 and 4 meters. We found that pointing to links in the Netscape window, that were displayed with a height of approximately 5 cm, was a difficult task and resulted in several misses before the link could actually be selected. This is due to the fact that the jitter of the vision system exceeds the height of the links. Another reason is the deictic nature of free hand pointing, which is not as accurate as mouse pointing. Nevertheless, pointing to images that where projected to 20 square cm worked well.

3.2 Menu selection

Menu selection is one of the major tasks when controlling an application via a WIMP interface. But, because of the findings in the last section and the difficulties experienced in reading menu names, the same metaphor can not be translated into the projection environment. To overcome this problem, we created large sensitive areas at the sides of the projection screen that bring up a menu when the user is pointing to that direction. To be useful, the menu must have a minimum extend, i.e. it should cover at least one quarter of the screen. In our example shown in figure 2a, the menu covers more than half the screen which induces a context switch for the user because the scene gets hidden. Furthermore, it also takes up a large amount of the physical environment the user is acting in.

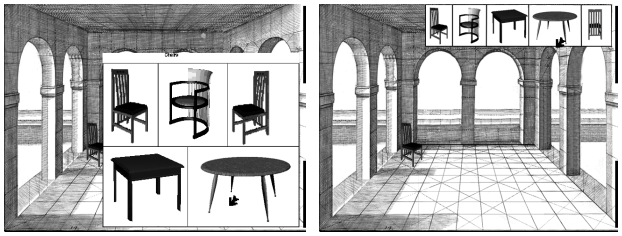

Figure 2: menu of chairs as a result of pointing to (a) the lower area at the right side of the screen, (b) the upper area

Figure 2b shows the best compromise between menu size for selection and still leaving the scene visible. Making the menu even smaller to display more items, for example, results in an unreadable menu. Another problem with menus is scale: where should all these sensitive areas be put if we want to choose between more than 10 different item groups? Therefore, we found this technique not to be applicable for our problem. A solution to this problem will be presented in the next section.

3.3 using real space

In the last section we explained the problem of context switching when using pop-up menus for selection purposes. One solution to overcome the restriction of not using menus is using real space as locations with inherent context. We define certain areas in real space as selection areas. Context switching is then performed by moving to that location. Similar to the example of retrieving data in real space [8], the user can move to the right edge of that area to retrieve a list of objects for selection and another when moving to the left side. By using this kind of context switching, the user is still aware of the current environment, i.e. still has „contact“ to the scene. Furthermore, by actively moving to a selection area the user „feels“ that he induces the context switch.

In addition to tracking the user, we also want to incorporate real objects that can be linked to virtual objects. Thereafter, the virtual objects will be moved according to the modification of the objects in real space. This integration of real objects allows the arrangement and discussion of a virtual set in real space.

4 The Prototype System

In this section we will give an overview of the prototype system, describe the image processing applied, and present a simple algorithm for estimating the pointing direction of the user. Thereafter, we will describe how the user interacts with the system and comment on the performance of the system.

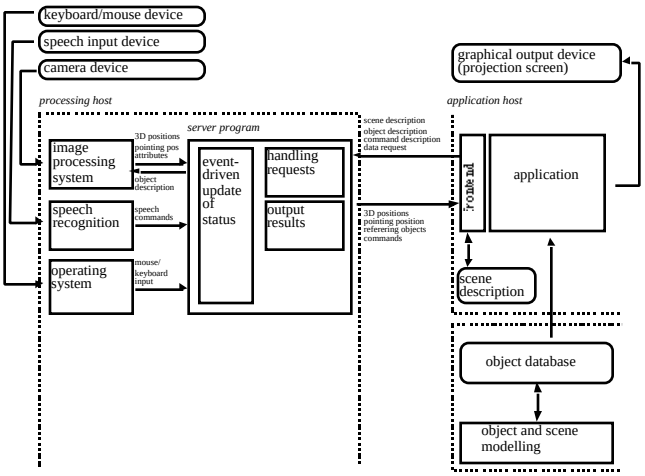

Figure 3: System architecture

4.1 System Overview

The system is divided up into the recognition part consisting of the image processing system and a speech recognition software, a server program, and the library front-end with the application program (see Figure 3). For speech recognition we use a PC-based system for recognizing simple commands [10]. The image processing system tracks the user via two video cameras. During initialization it receives a description of the objects to be tracked. Thereafter, it continuously sends 3D position data of the segmented objects and the users pointing-direction to the server program. The server program connects an application with the image processing, speech recognition and user input by mouse and keyboard. It updates the current states by an event driven loop. Upon request it will send data to the application continuously. The library front-end is a collection of methods that supply basic server communication as well as the basic interaction concepts for the Intuitive Interface. It reads a description of the objects to be tracked as well as a description of the virtual scene that will be displayed on the projection screen. This description is supplied by the application. It also links tracking results to actions and modifications in the virtual scene. The application program implements the methods to perform the desired film planning task. It uses the scene description and the library front-end to communicate with the server program. The application may run on hosts other than the server program.

4.2 Image Processing

The current version of the tracking system is based on color segmentation and blob analysis. The system tracks the body position by using a green marker that is attached to the user. Once this marker is segmented, a skin colored region is segmented within certain constraints on the right hand side of the user to obtain the position of the users hand. Currently, we segment the color image in uv color space by

using thresholds for u and v separately. The thresholds have been found by experimentation. Once the green marker is segmented, the right hand of the user is found relative to the center of this region. The search area for the skin colored hand region can be defined as a set of constraints. Simple heuristics together with a Kalman-filter result in a robust segmentation. This approach allows more than one person to be in the field of view of the cameras, as long as no other hand or head visually touch the search regions of the segmented regions. It also allows other skin colored objects to be around, as long as they do not lie in the constraint box of the user during initialization of the tracking.

After the marker and hand regions have been determined in each view of the scene, we apply a simple reconstruction scheme to determine the 3D position of the user and her hand. The reconstruction is based on known camera coordinates in the world coordinate system and camera parameters like angle of view and rotation angles that have been determined in a calibration phase.

4.3 Estimating pointing direction

The pointing direction is determined by using inverse kinematics applied to a simple model of the arm. Solving the inverse kinematic problem is not a simple task, because the underlying function is nonlinear and a unique mapping to solve the problem does not exist in general. However, since we only have a simple two-link structure, i.e., upper arm and lower arm, a simple closed-form solution can be derived [14]. In our case, shoulder and hand position do not provide a unique solution for the elbow position. However, by making assumptions about the upper arm angle γ in z-direction (lift of arm sideways), the problem of estimating the elbow position becomes two dimensional.

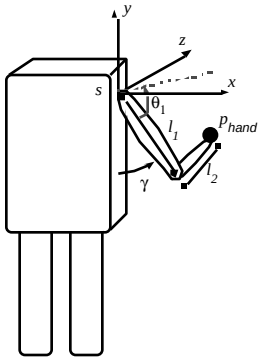

Figure 4: Simple arm model: upper and lower arm length (l_1, l_2) are predetermined, shoulder position (s) is determined relative to the body.

First, the hand position vector p_{hand} is put into the yz-plane by a rotation about the y-axis. Next, we solve for the links of the arm model: Given the position vector $X = (y, z)$ for a two-link structure, inverse kinematics solves for

$$\Theta = f^{-1}(X)$$

whereas $\Theta = (\theta_1, \theta_2)$ denote the angles of the two links. By applying elementary trigonometry the solution is:

$$\theta_1 = \tan^{-1}\left\{ \frac{-(l_2 \sin\theta_2) z + (l_1 + l_2 \cos\theta_2) y}{(l_2 \sin\theta_2) y + (l_1 + l_2 \cos\theta_2) z} \right\}$$

$$\theta_2 = \cos^{-1} \frac{(z^2 + y^2 - l_1^2 - l_2^2)}{2 l_1 l_2}$$

Thereafter, the upper arm vector $v = (0,0, l_1)$ is rotated about the x-, z-, and y-axis by θ_1, γ, and the predetermined x- and y-rotation values respectively. For determining γ we observed that the users arm is held closer to the body, i.e. about 175 degrees (assuming that pointing straight up at the ceiling, represents 0 degrees, and pointing straight down represents 180 degrees), and when pointing to the left side the arm is lifted up to 45 degrees, i.e. resulting in an angle of 130 degrees. While facing the cameras, the x-position of the hand relative to the shoulder position can be used to set the angle. Although this is a rather simple assumption, it leads to a correct pointing position and a more natural arm motion of the model. Not altering the angle results in a wrong pointing position (sometimes being off more than 100 cm) and a stiff looking arm motion of the model.

4.4 Interaction

Using the interaction paradigms of the Intuitive Interface described in section 3, we created part of an interactive film planing system.

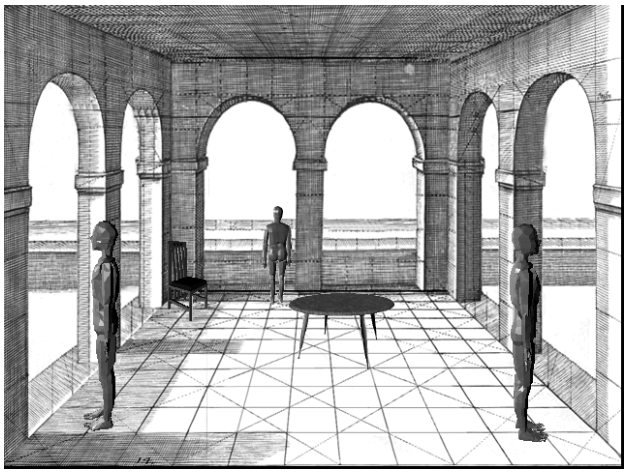

Figure 5: example scene of the prototype system

Currently, a generic scene can be arranged, i.e. objects can be moved around, rotated, or deleted, and new objects can be added to the scene. Figure 5 shows a configuration of

the example scene with 3 figures and 2 objects that have been arranged.

Selecting objects. We used two methods for selecting objects: a speech command and an auto-click operation when pointing to the same location for a while. The speech command worked well and is intuitive to use. However, while testing we often find it annoying to repeat the command again and again. The second approach works without a special command and is based on the fact that humans tend not to rest at the same location for a long time while pointing. By pointing to the same location for more than a second, the object becomes selected, i.e. the system locks onto the object. Releasing the object is performed in the same way. We found this technique easy to use and suggest that it should be combined with a speech command as an alternative.

Context switching. Context switching has to be performed when the user wants to add objects to the current scene. We use a selection area and combine this technique with a stage like metaphor, i.e. while the user is acting in the front part of the real space, the scene like figure 5 is shown on the screen. Here, objects may be moved around or manipulated. To add objects to the scene, the user moves backwards in real space. The scene zooms out, revealing a stage kind of setup that surrounds the previously shown scene (see figure 6). Here the user can select from various requisites like chairs, tables etc.

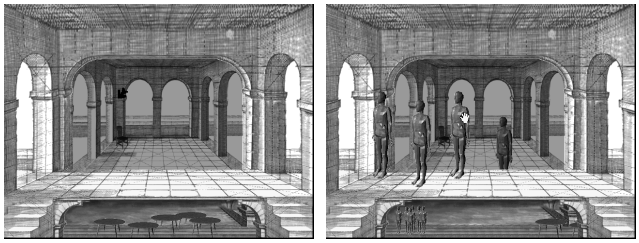

Figure 6: a) Stage setup for adding items to the scene, b) activated group shown on stage

To allow a greater selection, the objects are divided up into several groups shown on a conveyor belt at the bottom of the scene shown in figure 6. Each group is placed at a different room location in real space. The user can choose from these objects by moving to the specific location, which will move the conveyor belt to the appropriate position, and pointing to the group. Thereafter, the group of items will be moved on to the stage and the user is able to select an appropriate item (figure 6 b).

Linking real objects to virtual. In addition to tracking the user, we also track two wooden poles that are marked by colored signs to facilitate tracking. The poles can be linked to virtual objects by issuing a speech command. Thereafter, the virtual objects will be moved according to the modification of the poles (figure 7). This allows the arrangement to be evaluated in real space and a haptic feedback because of the weight of the poles.

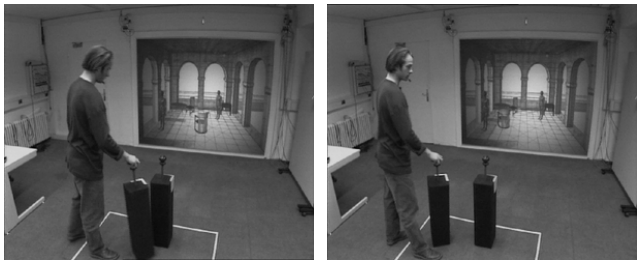

Figure 7: Using real objects to control virtual

4.5 Performance

The image processing system, as well as the speech recognition software and the film planning application program, currently run on a 200 MHz Pentium/Pro PC with two Matrox Meteor frame grabber boards at about 12 frames per second. The current prototype uses Macromedias Director for the film planning application. Using a second PC to run the application program, we get a frame rate of 20 fps for tracking the user, and a frame rate of about 16 fps for tracking the user and two wooden poles, if the poles are tracked on half the rate.

Table 1 shows the achieved accuracy of the tracking system grabbing half PAL-sized images in 4:2:2 format, i.e. resulting in a resolution of 192x144 pixels for the u- and v-channels respectively. The absolute accuracy is sufficient, because the visual feedback that is given by the screen cursor is far more important, similar to hand-eye feedback with mouse and mouse pointer. The accuracy has been measured at 30 different 3D positions from a distance of between 2.7 and 5 meters, table 1 shows the worst case. The pointing gesture has been tested while pointing to the projection screen (2 by 2.6 meters) from a distance of 3.2 meters.

table 1: Accuracy of the tracking system

absolute 3D-Position	9 cm
relative 3D-Position	2 cm
pointing (most cases)	10 cm
pointing (all cases)	20 cm
pointing jitter	15 cm

Most of the problems we had with our system currently stem from the jittering of the pointing position. This is due to the simple segmentation algorithm that cannot compensate for all shadowing situations and is sensible to hand movements of the user. We are currently working on an improvement of the segmentation using a lookup-table that is determined in a calibration phase. Another problem is the green marker the user has to wear: The user has to face the cameras all the time, which sometimes results in drop-outs of data when the marker is not visible to both cameras.

5 Empirical Study

First results of using the system showed that the use of real space and the memory-function are easy to learn. We found that the user must get used to the deictic gestures for moving objects around. Using real objects, on the other hand, was easy to use and faster than pointing and selecting. To get further insight of needs of film directors for such a planning tool, we made interviews with 6 directors working in different fields (feature film, documentary and experimental film). The age of the directors ranged between 31 and 57 years, and their prior contact to computer technology also varied. We chose the so called depth interviewing technique [7] which is a kind of conversation along some pre-determined questions. We first asked questions about the way the directors are used to work, demonstrated the prototype system thereafter, and, finally, asked about the personal impressions on using the system.

Most of the people asked did not like the fact that they had to move in real space for using the system. Specially, when adding objects to the scene, all persons mentioned that they would prefer a menu kind of selection on the screen. This indicates, that a discrete selection of objects should be preferred over influencing the process of selection using body movements. The reason for the directors mentioning „menus“, might stem from the (still) 2D looking scene displayed on the rear-projection and the studio or theater oriented view onto the virtual scene. User reactions might be different when using 3D scenes and camera movements. On the other hand, all persons said they liked the sensation of space that stems from the large rear-projection and the situation of the user in space.

Most work on set is done with actors. All directors agreed that such a planning tool would be good for planning the work with actors, specially in complex scenes, in conjunction with camera settings, movements and stage design. It would facilitate the discussion on particular camera movements, light setups, and set design, where often the intentions of the directors are misunderstood or hard to visualize using pen and pencil alone. Specially, for expendable feature films, a planning tool for complex scenes can reduce costs significantly. Using real objects to manipulate virtual objects was agreed to be intuitive and fast to use and, therefore, could be a good substitute to the commonly used models for planning a scene.

6 Conclusion

We presented a prototype system for intuitive film planning based on the concepts of the Intuitive Interface. The currently implemented system allows to arrange a virtual set. The environment makes strong use of space and body movements as well as deictic gestures to direct a computer generated scene. For estimating the pointing direction of the user we described the use of inverse kinematics in conjunction with a simple model of the arm. First results of an empirical study show that such a system would be applicable for planning complex scenes. The use of real space enhances the impression of space, though, directors would dislike to move in space to call functions. In the future, we plan to incorporate camera movements and improve the pointing operation. Thereafter, using space in conjunction with camera movements and real objects should be examined again in further usability studies.

References

[1] A. Azarbayejani, C. Wren and A. Pentland, *Real-Time 3-D Tracking of the Human Body*, Proceedings of IMAGE'COM 96, Bordeaux, France, May 1996.

[2] R. A. Bolt, *The Human Interface*, Lifetime Learning Publications, Belmont, California, 1984.

[3] U. Bröckl-Fox, *Real-Time 3D Interaction with up to 16 Degrees of Freedom from Monocular Video Image Flows*, International Workshop on Automatic Face- and Gesture-Recognition, June 26-28, Zürich, Switzerland, pp. 172-178, 1995.

[4] T. Darrell et. al., *A Novel Environment for Situated Vision and Behavior*, Proc. of CVPR-94 Workshop for Visual Behaviors, pp. 68-72, Seattle, Washington, June 1994.

[5] S. Feiner et. al., *Windows on the World: 2D Windows for 3D Augmented Reality*, Proc. UIST '93: ACM Symp. On User Interface Software and Technology, Atlanta GA, November 3-5, pp. 145-155, 1993.

[6] G. Fleischmann, M. Hoch and D. Schwabe, *FilmPlan: ein interaktives Filmplanungssystem*, Lab 1, Magazin der Kunsthochschule für Medien Köln, pp. 34-39, 1994.

[7] R.L. Gordon, *Interviewing. Strategy, Techiques and Tactics.* Homewood, Ill. 1969.

[8] M. Hoch, G. Fleischmann, *Social Environment: Towards an Intuitive User Interface*, 3D Image Analysis and Synthesis, Proceedings, November 18-19, Infix 1996, pp155-161

[9] M. Hoch, *Intuitive Schnittstelle*, Lab, Jahrbuch 1996/97 für Künste und Apparate, Verlag Walther König, Köln 1997.

[10] Lernout & Hauspie Speech Products, *User's Reference and Programming Guide*, Sint-Krispijnstraat 7, 8900 Ieper, Belgium, V3.010, July 1996.

[11] C. Maggioni, *GestureComputer - New Ways of Operating a Computer*, International Workshop on Automatic Face- and Gesture-Recognition, June 26-28, Zürich, Switzerland, pp 166-171, 1995.

[12] J.W. Sulivan, S.W. Tyler (Ed.), *Intelligent User Interfaces*, ACM press, New York 1991.

[13] M. C. Torrance, *Advances in Human-Computer Interaction: The Intelligent Room*, Working Notes of CHI '95 Research Symposium, May 6-7, Denver, Colorado, 1995.

[14] A. Watt, *Animating articulated structures*, in Advanced Animation and Reendering Techniques : theorie and practice, ACM Press, New York, 1992

[15] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, *Pfinder: Real-Time Tracking of the Human Body*, International Conference on Automatic Face and Gesture Recognition, Oct 96, Killington, Vermont, 1996, pp 51-56.

[16] F. A. Yates, *The Art of Memory*, Routledge & Kegan Paul/PLC, London, 1966.