Natural Interaction in Intelligent Spaces: Designing for Architecture and Entertainment

Flavia Sparacino
Sensing Places and MIT
flavia@sensingplaces.com
flavia@media.mit.edu

Keywords: Ambient Intelligence, Intelligent Architecture, Interactive Spaces, Interactive Entertainment

Abstract
The rapid evolution of computers' processing power, progress in projection and display technology, and their low cost, accompanied by recent advances in mathematical modeling, make available to space designers today sophisticated technologies which were once accessible only to research institutions or large companies. Thanks to wireless sensing techniques it is possible to endow a space with perceptual intelligence, and make it aware of how people use it, move in it, or react to it. Intelligent Spaces are relevant for several applications or tasks which range from surveillance to entertainment, from medical rehabilitation to artistic performance, from museum exhibit design to commerce. The author's work focuses on Narrative Spaces which are storytellers, able to articulate an informative or entertaining audio-visual narration for people interactively. Narrative Spaces communicate by use of large scale coordinated projections, sounds and displays whose contents are choreographed by the natural body movements or physical gestures of the people in them. This paper describes the guiding principles and modeling approaches that, according to the author, enable a robust modeling of user input and communication strategies for digital content presentation in Intelligent Narrative Spaces. It then provides examples of applications built according to the specified criteria.

1. Introduction
Designing responsive environments for various venues has become trendy today. Museums wish to create attractive "hands-on" exhibits that can engage and interest their visitors. Several research groups are building an "aware home" that can assist elderly people or chronic patients to have an autonomous life, while still calling for or providing immediate assistance when needed.

The design of these smart spaces needs to respond to several criteria. Their main feature is to allow people to freely move in them. Whether they are navigating a 3-D world or demanding assistance, we can't strap users with encumbering sensors and limiting tethers to make them interact with the space. Natural interaction, based on people's spontaneous gestures, movements and behaviors is an essential requirement of intelligent spaces. Capturing the user's natural input and triggering a corresponding action is however in many cases not sufficient to ensure the appropriate response by the system. We need to be able to interpret the users' actions in context and communicate to people information that is relevant to them, appropriate to the situation, and adequately articulated (simple or complex) at the right time.

On the basis of my work and research I will argue that intelligent spaces need to be supported by three forms of intelligence: *perceptual intelligence*, which captures people's presence and movement in the space in a natural and non-encumbering way; *interpretive intelligence*, which "understands" people's actions and is capable of making informed guesses about their behavior; and *narrative intelligence*, which presents us with information, articulated stories, images, and animations, in the right place, at the right time, all tailored to our needs and preferences.

All three forms of intelligence need to co-exist and co-operate for an Intelligent Space to be effective. Narrative intelligence is important so that the space provides us with relevant information that matches our interests and needs. We need systems able to select how much information to give, with what detail and composition, and how to sequence and articulate various fragments. A lack of interpretive intelligence will produce applications that are unable to determine the appropriate time to get our attention on a specific matter or story, and which nag the user about whether he or she wants this or that. Context modeling and behavior modeling are ways of approaching interpretive intelligence, the first one from the situation's perspective and the second one from the user's perspective. Perceptual Intelligence is about endowing spaces with eyes, ears, and sensors that allow them to capture how they are being used by people. This is in its full complexity still an open field of research: scene interpretation and object recognition, are for example active and open territories for scientific investigation.

Augmenting a space with all three forms of intelligence can be seen as endowing the space with a mind of its own, which transforms it from a simple container of people and things to an active participant and cooperating agent of our lives. Perceptual intelligence represents the bottom layer of this virtual brain, and processes sensorial inputs. Interpretive Intelligence is the middle layer whose role is to "make sense" of the input data: it identifies situations and people's behaviors. Narrative Intelligence is the upper layer, a bit like the brain cortex, and it regulates the output and communication between the intelligent space and us [Figure 1].

The above description of space intelligence has provided a high level framework for the author's research. In the remaining of this paper I will first illustrate in more detail the three forms of space intelligence. I will then describe incremental contributions to intelligence modeling for spaces I developed in the past years with a focus on applications for architecture and entertainment. The main contribution of this paper is to show that Bayesian networks are an ideal modeling tool for all three forms of space intelligence and provides a unifying mathematical description for robust sensing, user and context modeling, and articulated information delivery (storytelling).

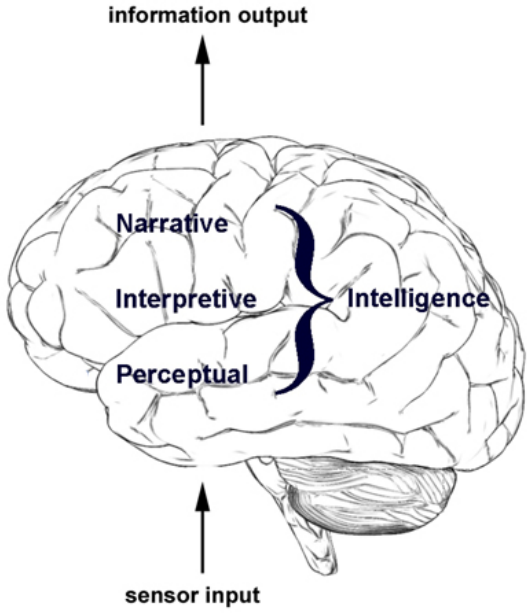


Figure 1. Layered Intelligence Model for Interactive Spaces

2. Related Work

The work here presented is highly interdisciplinary and it draws from various disciplines

2.1. Smart Spaces

The author began research in Smart Spaces and Natural Interfaces in collaboration with Alex Pentland at MIT [29] [33] [44] [45]. A variety of research groups today have taken this field of investigation further, with a main focus on assisted living and on creating spaces that are useful for ordinary everyday activities. Georgia Tech has developed the Aware Home, a place embedded with technology that enables older adults to age in place, and which helps people communicate with distant relatives and friends [22]. Microsoft's easy living is an intelligent environment that contains myriad devices that work together to provide users access to information and services [7]. These devices may be stationary, as with acoustic speakers or ceiling lights, or they may be mobile, as with laptop computers or mobile telephones. Microsoft's goal is to allow typical PC-focused activities to move off of a fixed desktop and into the environment [24]. The Stanford Interactive Workspaces project is exploring new possibilities for people to work together in technology-rich spaces with computing and interaction devices on many different scales [19]. Their work aims at augmenting a dedicated meeting space with large displays, wireless/multimodal I/O devices, and seamless integration of mobile and wireless appliances including handheld PCs. Rather than seeing the space as a pro-active facilitator Stanford's room is a reconfigurable platform for running a variety of applications. MIT's Intelligent Room is a highly interactive environment that uses embedded computation to observe and participate in normal,

everyday events. Microphone and camera arrays enable it to listen to people and observe what they do. People can speak with, gesture to, and interact with it in several ways [9] [6]. A robust agent-based software infrastructure supports the operation of these tools. The room's intelligence is modeled by a multi-agent society that builds a higher level and context based representation of the user's activity. Each activity has an associated software agent, called behavior agent, which responds to the user's actions and performs the corresponding appropriate reaction [12]. The room detects and reacts to activities such as watching a movie, meetings, entering the room. When new sensors are added to the space, the behavior rules need to be manually modified to take advantage of the new input information. The space intelligence model proposed in this paper uses instead a unified mathematical framework, Bayesian networks, both for perceptual intelligence and user modeling. Thanks to the properties of Bayesian networks, context or user behavior is modeled not as a set of fixed high level rules, but it is grounded on data observation, and therefore it can adapt or fine tune the room's response to the actual behavior of people in it, both instantly and through time. Additionally the Bayesian network architecture makes it easy to add new sensors or new responses to the room withought having to re-program the entire system to take new elements into account. As the applications of interest to the author are geared towards entertainment and to support new forms of interactive architecture, this paper also highlights the importance of and describes narrative intelligence, which allows the intelligent space to use its resources to articulate an audiovisual narration as a function of the user's estimated interests and behaviors in the space.

In a recent paper Emiliani and Stephanidis have examined the requirements that spaces with ambient-intelligence technologies need to have to support elderly people and people with disabilities According to their work, the main high-level design features of a system with ambient intelligence are that it be unobtrusive (i.e., many distributed devices are embedded in the environment, not intruding upon our consciousness unless we need them), personalized (i.e., it can recognize the user, and its behavior can be tailored to the user's needs), adaptive (i.e., its behavior can change in response to a person's actions and environment), and anticipatory (i.e., it anticipates a person's desires and environment as much as possible without mediation) [11]. The work presented in this paper uses similar requirements for an entertainment space (i.e. a museum) and it offers a layered model of intelligence with a unified mathematical representation which also includes narrative intelligence. Early work by the author on space intelligence applies some of the ideas discusses in this paper towards the design of an interactive museum exhibit [41].

2.2. Perceptual Intelligence and Natural Interaction

Pavlovic [27] and Wu [47] have explored use of hand gestures in human computer interaction, with an emphasis on 3-D tracking and multi-modal approaches for gesture recognition. Starner [42] has shown one of the first examples of effective gesture recognition using HMMs. Brand, Oliver, and Pentland [5] have demonstrated the higher performance of coupled HMMs for tasks which require gesturing with both hands at the same time. Campbell and others [8] have studied the effects of the appropriate feature choice for a gesture recognition task, using stereo vision.

The author has developed a variety of natural interfaces for art and entertainment installations [36] [37]. Her work on natural interfaces has grown from a monocular real time body tracking technique, called Pfinder (person finder), and the real time blob tracking approach associated with it and described in section 4.1.3 [46]. Stereo tracking of pointing and command gestures, and HMM based gesture recognition is discussed in [39]. Jojic, Brumitt, and Meyers [20] also use stereo cameras to detect people pointing and estimate the direction of their pointing. As opposed to the blob tracking approach developed by the author they use disparity maps which are less sensitive to lighting changes. In the blob-based approach, light invariance can be achieved using adaptation, or implementing color invariant classification [4].

2.3 Bayesian Networks for User Modeling and Interactive Narrative

The work of Pearl [28] is fundamental to the field of Bayesian networks. Jordan's book [21] had the merit of grouping together some of the major advancements since Pearl's 1988 book. Jensen [17][18] has written two thorough introductory books that provide a very good tutorial, or first reading, in the field. Bayesian networks have gained popularity in the early nineties, when they were successfully applied to medical diagnosis [13].

Albrecht et al [1], have been among the first to model the behavior of a participant in a computer game using Bayesian networks. They use several network structures to predict the user's current goal, next action, and next location in a multi-user Dungeon adventure game. With respect to their work the system here presented performs not just user state estimation, which can also be used to predict the next action that the user will do, but it also adds a probabilistic mechanism for content selection. The focus of the narrative intelligence model here presented is to function as a sensor-driven computational system that uses the same unifying mathematical framework (Bayesian Networks) for sensor modeling (tracking, gesture recognition), user estimation, and content selection. Jebara [16], uses CHMMS, which are a particular case of Dynamic

Bayesian Networks, to perform, first analysis, and then synthesis, of a player's behavior in a game. Conati et al. [10] have built an intelligent tutoring system able to perform knowledge assessment, plan recognition and prediction of students' actions during problem solving using Bayesian networks. Jameson [15], provides a useful overview of student modeling techniques, and compares the Bayesian network approach with other popular modeling techniques.

The narrative intelligence model here presented is inspired by work in probabilistic knowledge representation. Koller and Pfeffer have used probabilistic inference techniques that allow most frame bases knowledge representation systems to annotate their knowledge bases with probabilistic information, and to use that information to answer probabilistic queries. Their work is relevant to describe and organize content in any database system so that it can later be selected either by a typed probabilistic query or by a sensor driven query [23]. Using a content database, annotated probabilistically, the sto(ry)chastics system described in section 4.3. selects the most appropriate content segment at each time step, and it delivers, interactively in time and space, an audiovisual narration to the museum visitor as a function of the estimated visitor type

3. Criteria for Intelligent Space Design

3.1. Perceptual Intelligence

To transform a space in a computer and sensor-driven narrative space we need to be able to interact with the digital content on display in ways that are not as limiting as when we interact with the familiar keyboard and mouse. Furthermore, it is unrealistic to encumber participants with gloves, cables, or heavy virtual reality glasses: these fail to fully engage people as the technology dominates over the experience. Touch-based screens or hardware-based sensors tend to wear and break after use by hundreds of people and therefore require frequent replacement and a high maintenance load. Ideally we want to endow our interactive spaces with eyes, ears and perceptual intelligence so that they can interpret people's natural movement, gestures, and voice. Our spaces should be aware of visitors approaching an object of interest, of how they approach it, (speed, direction, pauses), they should be able to understand pointing gestures as commands, as well as occasionally understand more sophisticated gestures (zoom-in, zoom-out, rotate, move up, move down, me, you, etc.) and voice commands. Unencumbering computer vision interfaces or other wireless sensors, such as infrared sensors, electric field sensors, civil use sonars and radars, are the ideal input device or communication interface with the space.

The robustness and reliability of a natural interface is also very important. If the interface breaks often or does not work consistently, the "magic" of involvement and immersion in the interactive experience vanishes. Therefore ideally more than one sensor should be used to capture the participant's input. Cooperation of sensor modalities which have various degrees of redundancy and complementarity can guarantee robust, accurate perception. We can use the redundancy of the sensors to register the data they provide with one another. We then use the complementarity of the sensors to resolve ambiguity or to reduce errors when an environmental perturbation affects the system.

3.2. Interpretive Intelligence

To make good use of reliable measurements about the user, we need to be able to interpret our measurements in the context of what the user is trying to do with the digital media, or what we, as designers, want people to do so that they get the most out of the experiences we wish to offer. The same or similar gesture of the public can have different meanings according to the context and history of interaction. For example the same pointing gesture of the hand can be interpreted either as a zoom-in command gesture, or more simply, as a selection gesture. In a similar way, the system needs to develop expectations on the likelihood of the user's responses based on the specific context of interaction and content shown. These expectations influence in turn the interpretation of sensory data. Following on the previous example, rather than teaching both the user and the system to perform or recognize two slightly different gestures, one for zoom-in and one for selecting, we can simply teach the system how to correctly interpret slightly similar gestures, based on the context and history of interaction, by developing expectations on the probability of the follow-on gesture. In summary, our systems need to have a user model which characterizes the behavior and the likelihood of responses of the public. This model also needs to be flexible and should be adaptively revised by learning the user's interaction profile. Together with a user model the system should build a model of the "situation" in which the user is involved while interacting with digital media (context modeling).

3.3. Narrative Intelligence

In order to turn computers into articulated storytellers that respond to people's natural gestures and voice, we cannot simply model interaction as a list of coupled inputs and outputs. This simply defines a map of causes and effects that associates an action of the user to a response produced by the interactive space. Systems authored with this method tend to produce applications that are repetitive and shallow. We need instead narrative machines that are able to orchestrate stories whose composition and length can vary as a

function of the publics' interests. Just as a museum guide adapts his/her explanation of the artwork on display according to the visitors base knowledge and curiosity, our narrative engines should be able to take into account and adapt to the publics' needs. To accomplish this goal we need to model the story we wish to narrate so that it does takes into account and encompasses the user's intentions and the context of interaction. Consequently the story should develop on the basis of the system's constant evaluation of how the user's actions match the system's expectations about those actions, and the system's goals.

3.4. Intelligence Modeling

Over the last decade, a method of reasoning using probabilities, variously called belief networks, Bayesian networks, probabilistic causal networks, influence diagrams, knowledge maps, etc., has become popular within the AI community, and the machine learning, and pattern recognition communities [see section 2.3]. (Dynamic) Bayesian networks have been successfully applied to a variety of perceptual modeling tasks such as multimodal sensing for gesture recognition and sensor fusion [27], speech recognition [25], and body motion understanding [26]. Research in user and context modeling applies Bayesian networks to identify the behavior of a participant in a computer game [1], to interpret a car driver behavior [30], to understand the needs of a student [see section 2.3]. More recently the author has investigated Bayesian networks for story modeling and content selection in sensor-driven interactive narrative spaces [40]. In this paper I will argue that Bayesian networks are an ideal intelligence modeling tool as they can be used effectively to model respectively perceptual, interpretive, and narrative intelligence for interactive spaces.

4. Applications

This section describes three examples of spaces or space-components the author developed which each contribute, piece-wise, to the construction of intelligent environments. In some cases the system only has perceptual intelligence, whereas in others, the interpretive and narrative intelligence modeling is the focus of the contribution. As the aim of this paper is to provide a unified view of space modeling techniques, the following sections will describe the approach and results obtained. The reader will find a more accurate description of the implementation details in the included bibliography.

4.1. Perceptual Intelligence: Navigating the internet City

This section presents a natural interface to navigate inside a 3-D Internet city, using body gestures. This work uses a combination of computer vision and pattern recognition techniques to capture people's interaction in a natural way. A wide-baseline stereo pair of cameras is used to obtain 3-D body models of the user's hands and head. The interface feeds this information to an Hidden Markov Model (HMM) gesture classifier to reliably recognize the user's browsing commands. With regard to the intelligence modeling framework here described, Smyth [32] demonstrated that HMMs are equivalent to dynamic Bayesian networks To illustrate the features of this interface I describe its application to a custom built 3-D Internet browser which facilitates the recollection of information by organizing and embedding it inside a virtual city through which the user navigates [40].

4.1.1. Natural Interfaces: Motivation

Recent technological progress allows today most home users to be able to afford powerful graphics hardware and computer processors. With this equipment people can navigate in sophisticated 3-D graphical environments and play engaging computer games in highly realistic and fascinating 3-D landscapes. Such progress has not been paralleled by equivalent advances in man-machine interfaces to facilitate access and displacement in virtual worlds. People still use quite primitive and limiting interfaces: the joystick, button-activated game consoles, or the computer keyboard itself. Full immersion and skillful exploration of 3-D graphical environments is limited by the user's ability to use these interfaces, and the consequences of repetitive use often involve undesired and painful medical consequences to the user's wrists, fingers, arms, or shoulders. New, more natural interfaces are needed to navigate inside virtual worlds.

On the other hand people spend today an increasingly large amount of time exploring the Internet: a bi-dimensional environment, which is quite unsophisticated and simple compared to the previously mentioned popular computer games. While the Internet allows designers to author and display animated web pages, with moving text and images, the browsers we have available today are still quite primitive. Internet browsers are flat: they are based on the old multimedia metaphor of the book, with 2-D pages filled with links to other pages, and bookmarks as a memory aid, to represent and organize the information available on the net. The only advantage of such information-interface is its non-linearity and rapid, visible access to interconnected data. The main disadvantage is that it is easy to get disoriented while navigating the Internet, as we rapidly lose track of what we've seen before the current page, and do not have perspective of what is accessible from the current page. The Internet could benefit from the same 3-D graphical progress which has determined the surge of the computer games industry, and provide the public with 3-D graphical browsers to help us better visualize, organize, and remember information [34].

This section presents two connected contributions: an Internet 3-D web browser and a natural interface to browse it. Our browser is based on the architectural metaphor of the city and organizes information by embedding it inside a virtual urban space. Providing a natural interface to navigate in our 3-D web browser is similar to designing a new interface for a 3-D computer game. The author envisions that in a not so distant future we will have a large flat panel display in our home which is like a windows to the internet city. We will navigate through this world with natural gestures, and eventually also voice commands, in ways similar to the ones described in this document.

4.1.2. City of News: an Internet City in 3-D

City of News is an immersive, interactive web browser that makes use of people's strength at remembering the surrounding 3-D spatial layout. For instance, everyone can easily remember where most of the hundreds of objects in their house are located. We are also able to mentally reconstruct familiar places by use of landmarks, paths, and schematic overview mental maps. In comparison to our spatial memory, our ability to remember other sorts of information is greatly impoverished. City of News capitalizes on this ability by mapping the contents of URLs into a 3-D graphical world projected on a large screen. This gives a sense of the URLs existing in a surrounding 3-D environment and helps the user remember the previous exploration path leveraging off his/her spatial memory. The URLs are displayed so as to form an urban landscape of text and images through which people can navigate [Figures 2 and 3]. The 3-D web landscape is a city. Known cities' layout, architecture, and landmarks are input to the program and are used as orientation cues and organizing geometry. This virtual internet world grows dynamically as new information is loaded, and anchors our perceptual flow of data to a cognitive map of the virtual internet city. Following a link causes a new building to be raised in the district to which it belongs, conceptually, by the content it carries, and content to be attached onto its "façade". By mapping information to familiar places, which are virtually recreated, City of News stimulates in its users association of content to geography. The spatial, urban-like, distribution of information facilitates navigation of large information databases, like the Internet, by providing the user with a cognitive spatial map of data distribution.

To navigate this 3-D environment, people sit or stand in front of a large screen [Figures 4 and 5] and use hand gestures to explore or load new data. Side-pointing gestures allow users to navigate along an information path. Both arms up drive the virtual camera above the City and give an overall color-coded view of the urban information distribution. When a new building is raised and the corresponding content is loaded, the virtual camera will automatically move to a new position which corresponds to an ideal viewpoint for the information landscape. All the virtual camera movements are smooth interpolations between "camera anchors" that are invisibly dynamically loaded in the space as it grows. These anchors are like rail tracks which provide optimal viewpoints and constrain navigation so that the user is never lost in the virtual world.

Figure 2. Aerial view of City of News

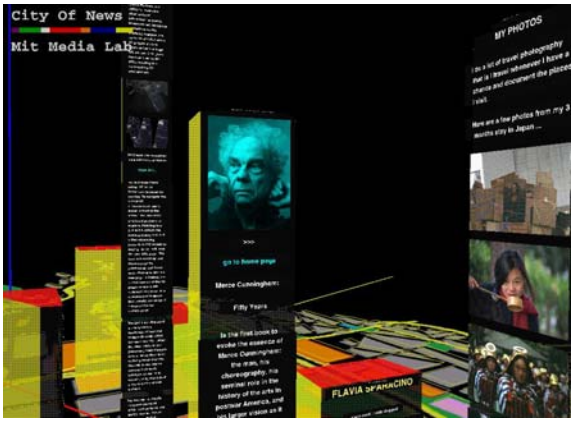


Figure 3. City of News after exploration

Figure 4. User standing at the Interactive Setup

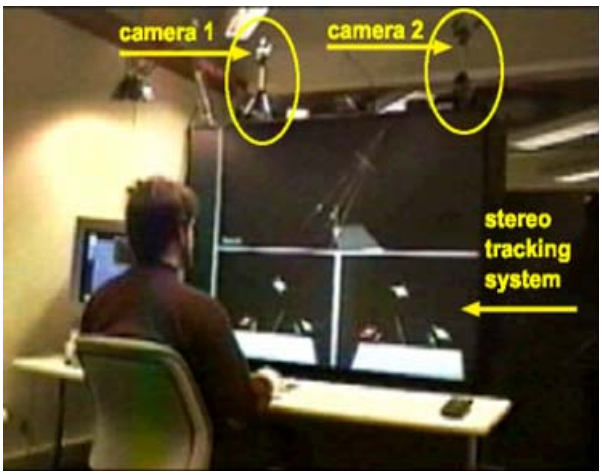


Figure 5. User sitting at the Interactive Setup

4.1.3. 2-D Blob Tracking

Real-time 3-D tracking is a method for estimation of 3-D geometry from blob features. The notion of "blobs" as a representation for image features has a long history in computer vision. The term "blob" is somewhat self-explanatory ("something of vague or indefinite form"), but a useful definition from a computational point of view is that a blob is defined by some visual property that is shared by all the pixels contained in the blob and is not shared by surrounding pixels. This property could be color, texture, brightness, motion, shading, a combination of these, or any other salient spatio-temporal property derived from the signal (the image sequence). Blobs are usually thought of as regions with dimensions that are roughly the same order-of-magnitude, in part because we have special terms for other features, e.g., "contours", "lines", or "points". But these other features can also be viewed as degenerate cases of blobs, and, in fact, straight contours and points are perfectly well represented by the blob model.

Blobs can be reliably tracked even in complex, dynamic scenes, and that they can be extracted in real-time without the need for special purpose hardware. These properties are particularly important in applications that require tracking people, and for this reason they have been used for real-time whole-body human interfaces [46] and for real-time recognition of American Sign Language hand gestures [42]. In the City of News setup, in which the upper body is used as the navigating interface to the Internet browser, it is important to have a more exact knowledge of body-parts position. A monocular system would not be able to accurately recover the location pointed at by the user in the 3-D landscape. A projection error onto the 3-D landscape would cause the user to navigate to a completely different location than what he/she intended. Not having such precision available would be equivalent to having a mouse with a coarse resolution which can cause a user to click and launch undesired applications, or involuntarily click on a different link than the desired one. A stereo tracking system, with the ability to recover the 3-D geometry of the user's input features – hands and head – is an important step towards precise and reliable man-machine interfaces to explore 3-D data.

For both 2-D and 3-D blobs, there is a useful physical interpretation of the blob parameters. The mean represents the geometric center of the blob area (2-D) or volume (3-D). The covariance, being symmetric, can be diagonalized via an eigenvalue decomposition to yield a rotation matrix and a diagonal size matrix. The diagonal size matrix represents the size of the blob along independent orthogonal object-centered axes and the rotation matrix brings this object-centered basis in alignment with the world coordinate basis. This decomposition and physical interpretation is important for estimation, because the shape is constant (or slowly varying) while the rotation is dynamic.

4.1.4. Person Tracking and Shape Recovery

The useful 3-D geometry of the human body is estimated from blob correspondences in displaced cameras. The relevant unknown 3-D geometry includes the shapes and motion of 3-D objects, and optionally the relative orientation of the cameras and the internal camera geometries. The goal is to recover the 3-D shape from the 2-D shapes [Figure 6].

The system first uses a color classifier to identify pixels with a skin color by log likelihood calculation in the MAP sense, as described in [47]. It then determines all connected skin-colored regions in the image with the k-means algorithm. In then discards the smaller blobs until three blobs are found: left hand, right hand, and head. Simple heuristics allow the program to easily assign which blobs correspond to which body part: the

head is usually on top and in the center, and at start the right hand is usually on the right-hand side of the head. When a person first enters the space, the stereo calibration is obtained recursively by incorporating the three blob correspondences (face, left hand, right hand) into a EKF/LM estimator [2][3]. The calibration parameters converge typically in the first 20 to 40 frames (roughly 2 to 4 seconds), if there is enough motion; longer if there is little motion. To evaluate the calibration quantitatively, the right hand acts as a 3-D pointer and traces the 3-D shape of known objects. The error of reconstruction of a hand position is on the order of 2 to 3cm when the user is up to 4 meters away from the screen. This error is due both to estimation error and the crudeness of using the hand position to represent a point in space.

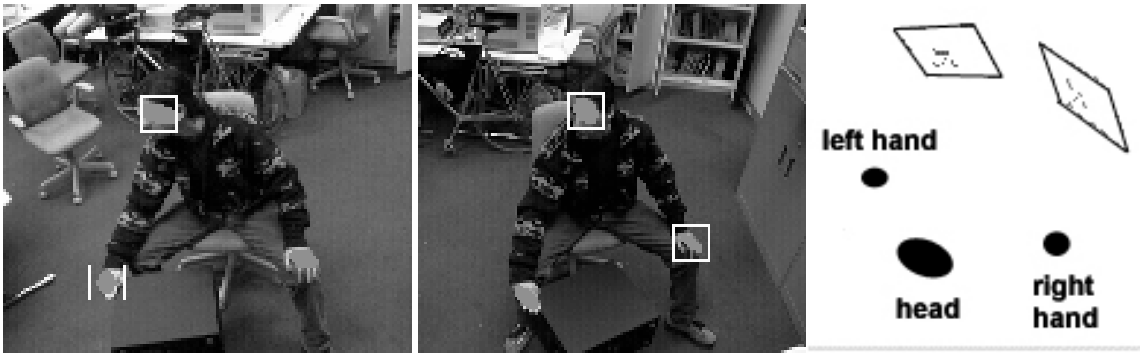


Figure 6. 3-D tracking of user's hands and head, view from the left and right camera

4.1.5. Gesture Recognition

A gesture-based interface mapping interposes a layer of pattern recognition between the input features and the application control. When an application has a discrete control space, this mapping allows patterns in feature space, better known as gestures, to be mapped to the discrete inputs. The set of patterns form a gesture-language that the user must learn. To navigate through the Internet 3-D city the user stands in front of the screen and uses hand gestures. All gestures start from a rest position given by the two hands on the table in front of the body. Recognized command gestures are [figures 8 and 9]:

- "follow link" → "point-at-correspondent-location-on-screen"
- "go to previous location" → "point left"
- "go to next location" → "point right"
- "navigate up" → "move one hand up"
- "navigate down" → "move hands toward body"
- "show aerial view" → "move both hands up"

Gesture recognition is accomplished by HMM modeling of the navigating gestures [31] [Figure 7]. The feature vector includes velocity and position of hands and head, and blobs' shape and orientation. We use four states HMMs with two intermediate states plus the initial and final states. Entropic's Hidden Markov Model Toolkit (HTK: http://htk.eng.cam.ac.uk/) is used for training [48]. For recognition we use a real-time C++ Viterbi recognizer.

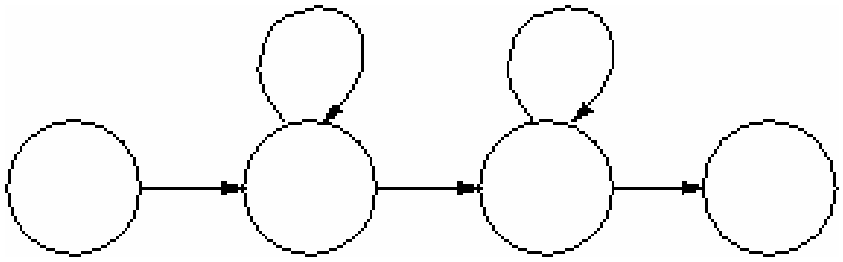


Figure 7. Four state HMM used for gesture recognition

Figure 8. Navigating gestures in City of News (user sitting)

Figure 9. Navigating gestures in City of News at SIGGRAPH 2003 (user standing)

4.1.6. Comments

I described an example of space, or space corner, which could be in the living room in our home, or in the lobby of a museum, in which perceptual intelligence, modeled by computer vision and Hidden Markov Models, a particular case of a Bayesian Networks, provides the means for people to interact with a 3-D world in a natural way. This is only a first step towards intelligence modeling. Typically an intelligent space would have a variety of sensors to perceive our actions in it: visual, auditory, temperature, distance range, etc. Multimodal interaction and sensor fusion will be addressed in future developments of this work.

4.2. Interpretive Intelligence: Modeling user preferences in the museum space

This section addresses interpretive intelligence modeling from the user's perspective. The chosen setting is the museum space, and the goal is to identify people's interests based on how they behave in the space.

4.2.1. User Modeling: Motivation

In the last decade museums have been drawn into the orbit of the leisure industry and compete with other popular entertainment venues, such as cinemas or the theater, to attract families, tourists, children, students, specialists, or passerbiers in search of alternative and instructive entertaining experiences. Some

people may go to the museum for mere curiosity, whereas others may be driven by the desire of a cultural experience. The museum visit can be an occasion for a social outing, or become an opportunity to meet new friends. While it is not possible to design an exhibit for all these categories of visitors, it is desirable for museums to attract as many people as possible. Technology today can offer exhibit designers and curators new ways to communicate more efficiently with their public, and to personalize the visit according to people's desires and expectations [35].

When walking through a museum there are so many different stories we could be told. Some of these are biographical about the author of an artwork, some are historical and allow us to comprehend the style or origin of the work, and some are specific about the artwork itself, in relationship with other artistic movements. Museums usually have large web sites with multiple links to text, photographs, and movie clips to describe their exhibits. Yet it would take hours for a visitor to explore all the information in a kiosk, to view the VHS cassette tape associated to the exhibit and read the accompanying catalogue. Most people do not have the time to devote or motivation to assimilate this type of information, and therefore the visit to a museum is often remembered as a collage of first impressions produced by the prominent features of the exhibits, and the learning opportunity is missed. How can we tailor content to the visitor in a museum so as to enrich both his learning and entertaining experience ? We want a system which can be personalized to be able to dynamically create and update paths through a large database of content and deliver to the user in real time during the visit all the information he/she desires. If the visitor spends a lot of time looking at a Monet, the system needs to infer that the user likes Monet and should update the narrative to take that into account. This research proposes a user modeling method and a device called the museum wearable to turn this scenario into reality.

4.2.2. The Museum Wearable

Wearable computers have raised to the attention of technological and scientific investigation [43] and offer an opportunity to "augment" the visitor and his perception/memory/experience of the exhibit in a personalized way. The museum wearable is a wearable computer which orchestrates an audiovisual narration as a function of the visitor's interests gathered from his/her physical path in the museum and length of stops. It offers a new type of entertaining and informative museum experience, more similar to mobile immersive cinema than to the traditional museum experience [figure 10].

The museum wearable [38] is made by a lightweight CPU hosted inside a small shoulder pack and a small, lightweight private-eye display. The display is a commercial monocular, VGA-resolution, color, clip-on screen attached to a pair of sturdy headphones. When wearing the display, after a few seconds of adaptation, the user's brain assembles the real world's image, seen by the unencumbered eye, with the display's image seen by the other eye, into a fused augmented reality image.

Figure 10. The museum wearable used by museum visitors

The wearable relies on a custom-designed long-range infrared location-identification sensor to gather information on where and how long the visitor stops in the museum galleries. A custom system had to be built for this project to overcome limitations of commercially available infrared location identification systems such as short range and narrow cone of emission. The location system is made by a network of small infrared devices, which transmit a location identification code to the receiver worn by the user and attached to the display glasses. [38]

The museum wearable plays out an interactive audiovisual documentary about the displayed artwork on the private-eye display. Each mini-documentary is made by small segments which vary in size from twenty

seconds to one and half minute. A video server, written in C++ and DirectX-8, plays these assembled clips and receives TCP/IP messages from another program containing the information measured by the location ID sensors. This server-client architecture allows the programmer to easily add other client programs to the application, which communicate to the server information from other possible sources, such as sensors or cameras placed along the museum aisles, which measure the crowdedness of the galleries or, upon an explicit request, affinities with other visitors using the same device. The client program reads IR data from the serial port, and the server program does inference, content selection, and content display [Figure 11].

The ongoing robotics exhibit at the MIT Museum provided an excellent platform for experimentation and testing with the museum wearable [figures 12]. This exhibit, called Robots and Beyond, and curated by Janis Sacco and Beryl Rosenthal, features landmarks of MIT's contribution to the field of robotics and Artificial Intelligence. The exhibit is organized in five sections: Introduction, Sensing, Moving, Socializing, and Reasoning and Learning, each including robots, a video station, and posters with text and photographs which narrate the history of robotics at MIT. There is also a large general purpose video station with large benches for people to have a seated stop and watch a PBS documentary featuring robotics research from various academic institutions in the country.

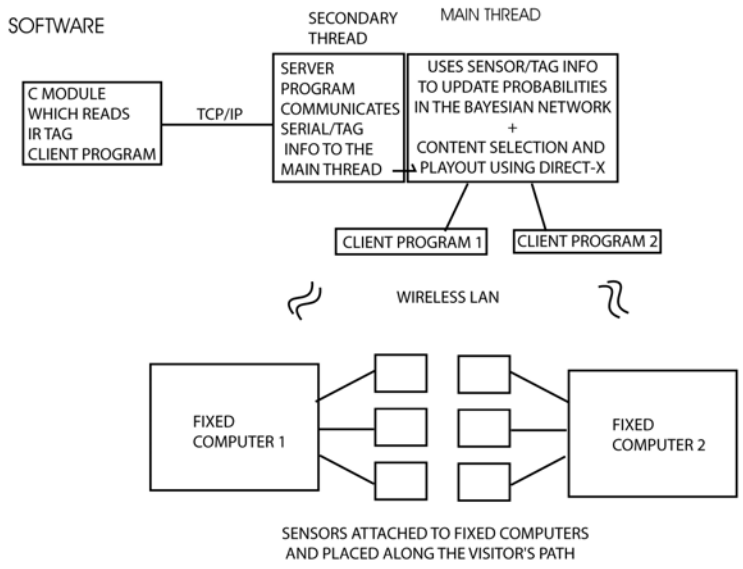


Figure 11. Software architecture of the museum wearable **Figure 12.The MIT robotics exhibit**

4.2.3. Sensor-driven understanding of visitors' interests with Bayesian networks

In order to deliver a dynamically changing and personalized content presentation with the museum wearable a new content authoring technique had to be designed and implemented. This called for an alternative method than the traditional complex centralized interactive entertainment systems which simply read sensor inputs and map them to actions on the screen. Interactive storytelling with such one-to-one mappings leads to complicated control programs which have to do an accounting of all the available content, where it is located on the display, and what needs to happen when/if/unless. These systems rigidly define the interaction modality with the public, as a consequence of their internal architecture, and lead to presentations which have shallow depth of content, are hard to modify, ad hoc, and prone to error. The main problem with such content authoring approaches is that they acquire high complexity when drawing content from a large database, and once built, they are hard to modify or to expand upon. In addition, when they are sensor-driven they become depended on the noisy sensor measurements, which can lead to errors and misinterpretation of the user input. Rather than directly mapping inputs to outputs, the system should be able to "understand the user" and to produce an output based on the interpretation of the user's intention in context.

In accordance with the simplified museum visitor typology discussed in [38] the museum wearable identifies three main visitor types: the busy, selective, and greedy visitor type. The greedy type, wants to know and see as much as possible, and does not have a time constraint, the busy type just wants to get an overview

of the principal items in the exhibit, and see little of everything, and the selective type, wants to see and know in depth only about a few preferred items. The identification of other visitor types or subtypes has been postponed to future improvements and developments of this research. The visitor type estimation is obtained probabilistically with a Bayesian network using as input the information provided by the location identification sensors on where and how long the visitor stops, as if the system was an invisible storyteller following the visitor in the galleries and trying to guess his preferences based on the observation of his/her external behavior.

The system uses a Bayesian network to estimate the user's preferences taking the location identification sensor data as the input or observations of the network. The user model is progressively refined as the visitor progresses along the museum galleries: the model is more accurate as it gathers more observations about the user. Figure 13 shows the Bayesian network for visitor estimation, limited to three museum objects (so that the figure can fit in the document), selected from a variety of possible models designed and evaluated for this research.

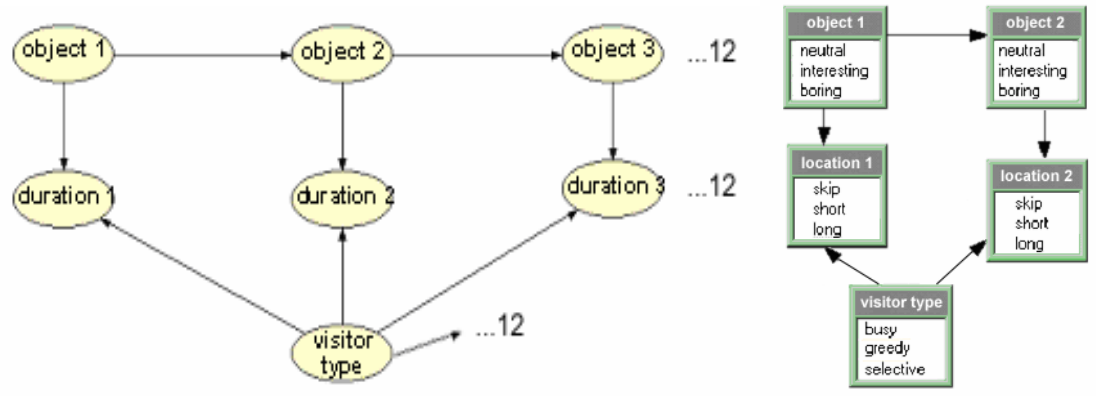


Figure 13. Chosen Bayesian Network model to estimate the visitor type

4.2.4. Model Description, Learning and Validation

In order to set the initial values of the parameters of the Bayesian network, experimental data was gathered on the visitors' behavior at the Robots and Beyond exhibit. According to the VSA (Visitor Studies Association, http://museum.cl.msu.edu/vsa), timing and tracking observations of visitors are often used to provide an objective and quantitative account of how visitors behave and react to exhibition components. This type of observational data suggests the range of visitor behaviors occurring in an exhibition, and indicates which components attract, as well as hold, visitors' attention (in the case of a complete exhibit evaluation this data is usually accompanied by interviews with visitors, before and after the visit). During the course of several days a team of collaborators tracked and make annotations about the visitors at the MIT Museum. Each member of the tracking team had a map and a stop watch. Their task was to draw on the map the path of individual visitors, and annotate the locations at which visitors stopped, the object they were observing, and how long they would stop for. In addition to the tracking information, the team of evaluators was asked to assign a label to the overall behavior of the visitor, according to the three visitor categories earlier described: "busy", "greedy", and "selective" [figure 13].

A subset of twelve representative objects of the Robots and Beyond exhibit, were selected to evaluate this research, to shorten editing time [Figure 14]. The geography of the exhibit needs to be reflected into the topology of the network, as shown in figure 15. Additional objects/nodes of the modeling network can be added later for an actual large scale installation and further revisions of this research.

The visitor tracking data is used to learn the parameters of the Bayesian network. The model can later be refined, that is the parameters can be fine tuned, as more visitors experience the exhibit with the museum wearable. The network has been tested and validated on this observed visitor tracking data by parameter learning using the Expectation Maximization (EM) algorithm, and by performance analysis of the model with the learned parameters, with a recognition rate of 0.987. More detail can be found in: Sparacino, 2003.

Figures 16, 17 and 18 show state values for the network after two time steps To test the model, I introduced evidence on the duration nodes, thereby simulating its functioning during the museum visit. The reader can verify that the system gives plausible estimates of the visitor type, based on the evidence introduced in the system. The posterior probabilities in this and the subsequent models are calculated using Hugin,

(www.hugin.com) which implements the Distribute Evidence and Collect Evidence message passing algorithms on the junction tree.

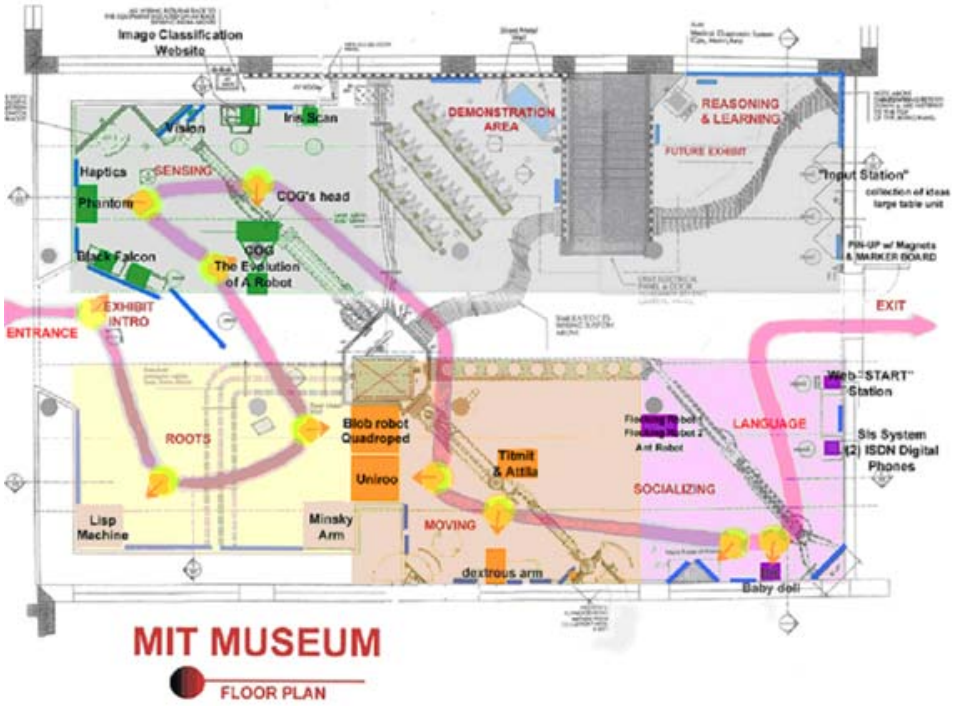


Figure 14. Chosen Bayesian Network model to estimate the visitor type

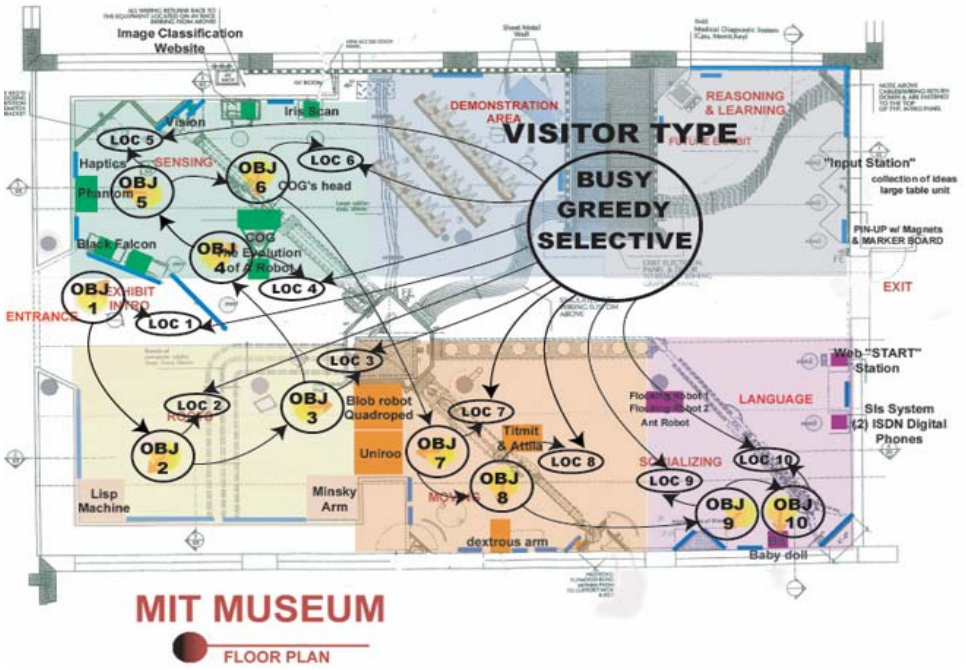
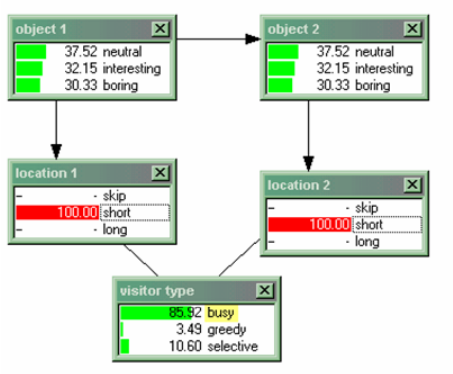


Figure 15. Chosen Bayesian Network model to estimate the visitor type



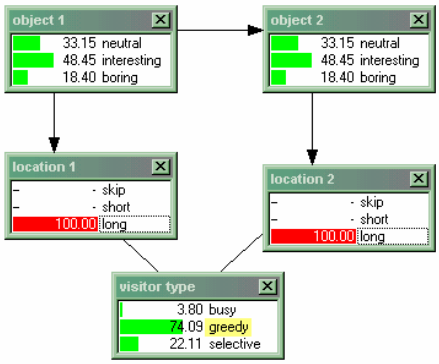
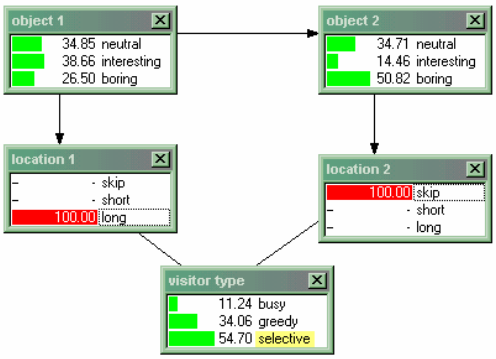


Figure 16. Test case 1. The visitor spends a short time both with the first and second object –> the network gives the highest probability to the busy type (0.8592)

Figure 17. Test case 2. The visitor spends a long time both with the first and second object –> the network gives the highest probability to the greedy type (0.7409)



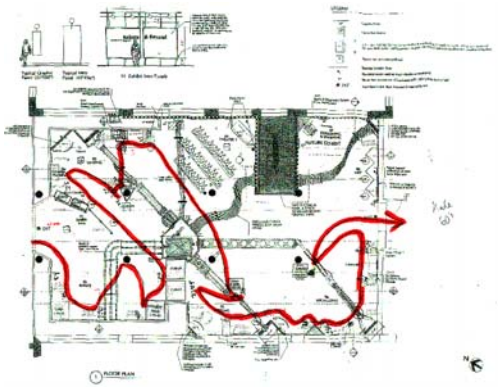


Figure 18. Test case 3. The visitor spends a long time with the first object and skips the second object –> the network gives the highest probability to the selective type (0.5470)

Figure 19. Tracking data sheet for a visitor's path in the museum space

4.2.5. Comments

Identifying people's preferences and typologies is relevant not only for museums but also in other domains such as remote healthcare, new entertainment venues, or surveillance. Various approaches to user modeling have been proposed in the literature. The advantage of the Bayesian network modeling here described is that it can be easily integrated in a multi-layer framework of space intelligence in which both the bottom perceptive layer and the top narrative layer are also modeled with the same technique. Therefore, as described above, both sensing and user typology identification can be grounded on data and can easily adapt to the behavior of people in the space. This work does not explicitly address situation modeling, which is an important element of interpretive intelligence, and which is the objective of future developments of this research.

4.3. Narrative Intelligence: Sto(ry)chastics

This section presents sto(ry)chastics, a user-centered approach for computational storytelling for real-time sensor-driven multimedia audiovisual stories, such as those that are triggered by the body in motion in a sensor-instrumented interactive narrative space. With sto(ry)chastics the coarse and noisy sensor inputs are coupled to digital media outputs via a user model (see previous section), which is estimated probabilistically by a Bayesian network [40]

4.3.1. Narrative Intelligence: Motivation

Sto(ry)chastics, is a first step in the direction of having suitable authoring techniques for sensor-driven interactive narrative spaces. It allows the interactive experience designer to have flexible story models, decomposed in atomic or elementary units, which can be recombined into meaningful sequences at need in

the course of interaction. It models both the noise intrinsic in interpreting the user's intentions as well as the noise intrinsic in telling a story. We as humans do not tell the same story in the same way all the time, and we naturally tend to adapt and modify our stories to the age/interest/role of the listener. This research also shows that, Bayesian networks are a powerful mathematical tool to model noisy sensors, noisy interpretation of intention, and noisy stories.

4.3.2. Editing Stories for Different Visitor Types and Profiles

Sto(ry)chastics works in two steps. The first is user type estimation as described in the previous section. The next step is to assemble a mini-story for the visitor, relative to the object he/she is next to. Most of the audio-visual material available for art and science documentaries tends to fall under a set of characterizing topics. After an overview of the audio-visual material available at MIT's Robots and Beyond exhibit, the following content labels, or bins were identified to classify the component video clips:

- Description of the artwork: what it is, when it was created (answers: when, where, what)
- Biography of author: anecdotes, important people in artist's life (answers: who)
- History of the artwork: previous relevant work of the artist
- Context: historical, what is happening in the world at the time of creation
- Process: particular techniques used or invented to create the artwork (answers: how)
- Principle: philosophy or school of thought the author believes in when creating the artwork (answers: why)
- Form and Function: relevant style, form and function which contribute to explain the artwork.
- Relationships: how is the artwork related to other artwork on display
- Impact: the critics' and the public's reaction to the artwork

This project required a great amount of editing to be done by hand (non automatically) in order to segment the two hours of video material available for the exhibit in the smallest possible complete segments. After this phase, all the component video clips were given a name, their length in seconds was recorded into the system, and they were also classified according to the list of bins described above. The classification was done probabilistically, that is each clip has been assigned a probability (a value between zero and one) of belonging to a story category. The sum of such probabilities for each clip needs to be one. The result of the clip classification procedure, for a subset of available clips, is shown in Table 8.

To perform content selection, conditioned on the knowledge of the visitor type, the system needs to be given a list of available clips, and the criteria for selection. There are two competing criteria: one is given by the total length of the edited story for each object, and the other is given by the ordering of the selected clips. The order of story segments guarantees that the curator's message is correctly passed on to the visitor, and that the story is a "good story", in that it respects basic cause-effect relationships and makes sense to humans. Therefore the Bayesian network described in the earlier needs to be extended with additional nodes for content selection [Figures 20 and 21]. The additional "good story" node, encodes, as prior probabilities, the curator's preferences about how the story for each object should be told. To reflect these observations the Bayesian network is extended to be an influence diagram [14]: it will include decision nodes, and utility nodes which guide decisions. The decision node contains a list of all available content (movie clips) for each object. The utility nodes encode the two selection criteria: length and order. The utility node which describes length, contains the actual length in seconds for each clip. The length is transcribed in the network as a positive number, when conditioned on a preference for long clips (greedy and selective types). It is instead a negative length if conditioned on a preference for short content segments (busy type). This is because a utility node will always try to maximize the utility, and therefore length is penalizing in the case of a preference for short content segments. The utility node which describes order, contains the profiling of each clip into the story bins described earlier, times a multiplication constant used to establish a balance of power between "length" and "order". Basically order here means a ranking of clips based on how closely they match the curator's preferences expressed in the "good story" node. By means of probability update, the Bayesian network comes up with a "compromise" between length and order and provides a final ranking of the available content segments in the order in which they should be played.

Sto(ry)chastics is adaptive in two ways: it adapts both to individual users and to the ensemble of visitors of a particular exhibit. For individuals, even if the visitor exhibits an initial "greedy" behavior, it can later adapt to the visitor's change of behavior. It is important to notice that, reasonably and appropriately, the system "changes its mind" about the user type with some inertia: i.e. it will initially lower the probability for a greedy type until other types gain probability. Sto(ry)chastics can also adapt to the collective body of its users. If a count of busy/greedy/selective visitors is kept for the exhibit, these numbers can later become priors of the corresponding nodes of the network, thereby causing the entire exhibit to adapt to the collective body of its

users through time. This feature can be seen as “collective intelligence” for a space which can adapt not just to the individual visitors but also to the set of its visitors.

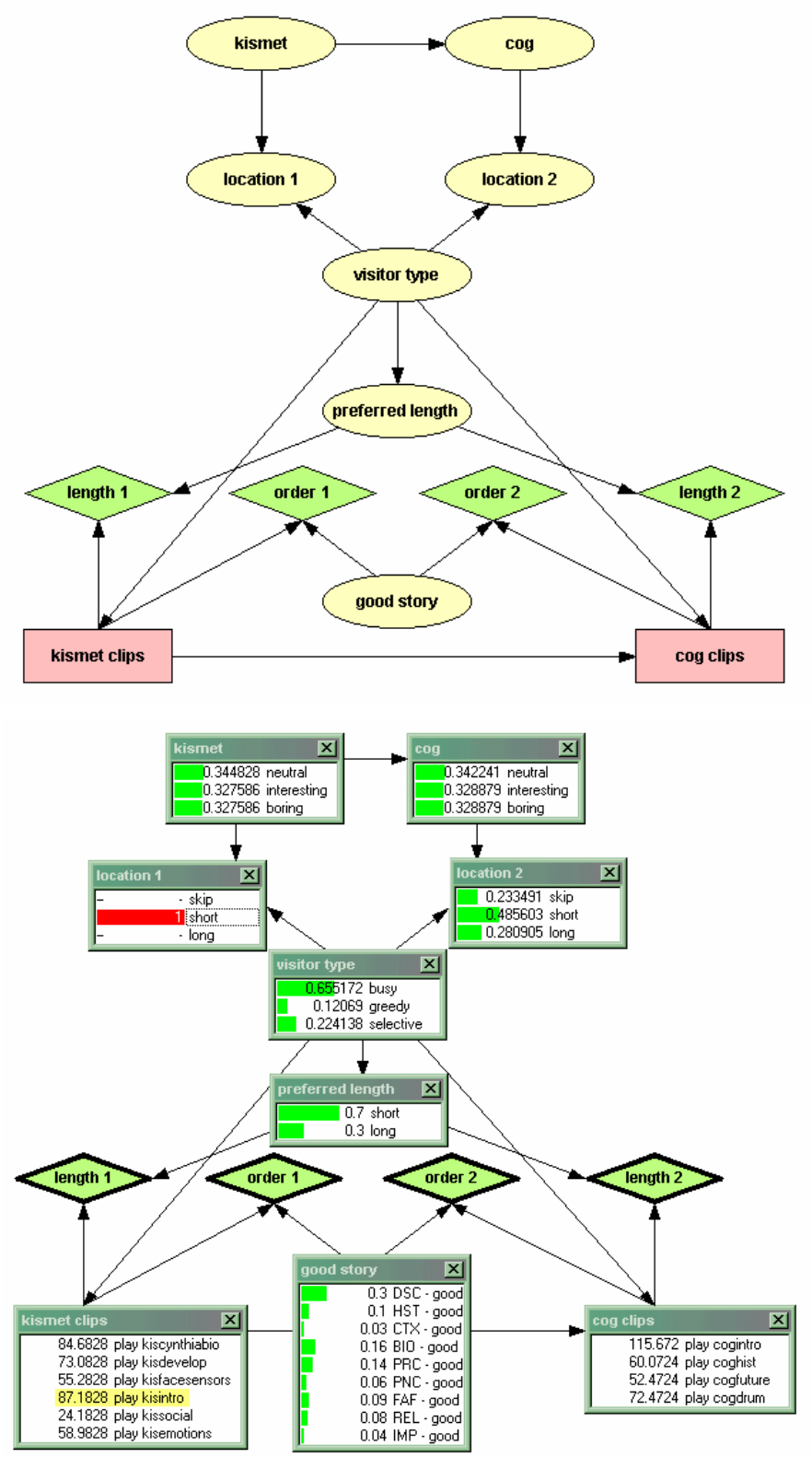


Figure 20. Extension of the sto(ry)chastics Bayesian network to perform content selection

4.3.3. Comments

The main contribution of this section is to show that (dynamic) Bayesian networks are a powerful modeling technique to couple inputs to outputs for real time sensor-driven multimedia audiovisual stories, such those that are triggered by the body in motion in a sensor-instrumented interactive narrative space. Sto(ry)chastics has implications both for the human author (designer/curator) which is given a flexible modeling tool to organize, select, and deliver the story material, as well as the audience, that receives personalized content only when and where it is appropriate. Sto(ry)chastics proposes an alternative to complex centralized interactive entertainment programs which simply read sensor inputs and map them to actions on the screen. These systems rigidly define the interaction modality with the public, as a consequence of their internal architecture. Sto(ry)chastics delivers an audiovisual narration to the visitor as a function of the estimated type, interactively in time and space. The model has been tested and validated on observed visitor tracking data using the EM algorithm. The interpretation of sensor data is robust in the sense that it is probabilistically weighted by the history of interaction of the participant as well as the nodes which represent context. Therefore noisy sensor data, triggered for example by external or unpredictable sources, is not likely to cause the system to produce a response which does not "make sense" to the user.

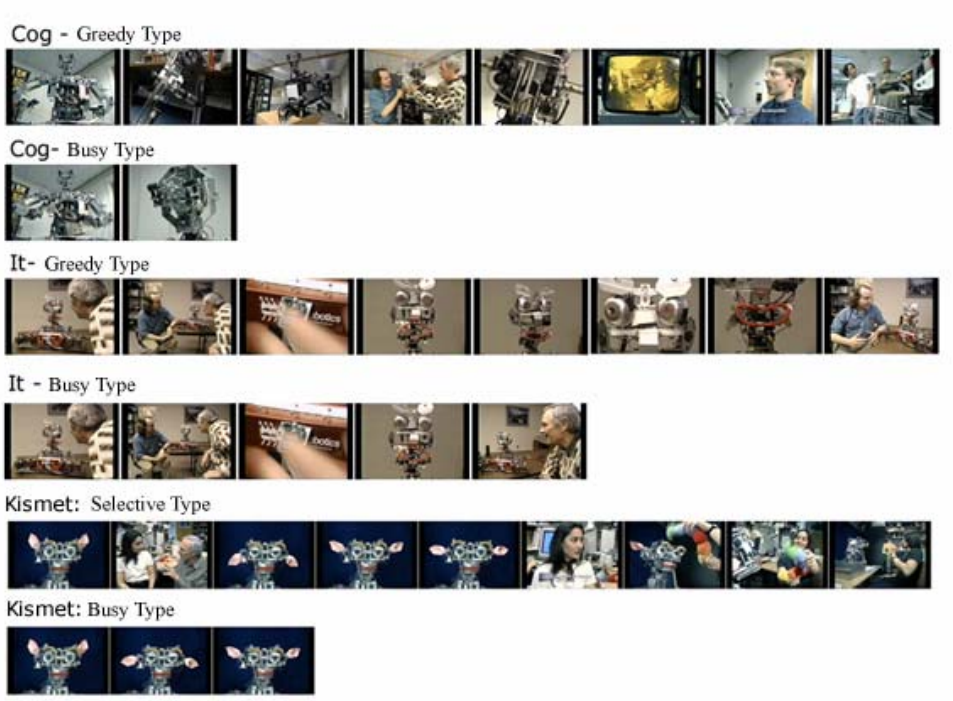


Figure 21. Storyboards from various video clips shown on the museum wearable's display at MIT Museum's Robots and Beyond Exhibit

5. Discussion and Conclusions

This paper presented a layered architecture of space intelligence, which the author believes is necessary to design body-driven interactive narrative spaces with robust sensing, tailored to the users' needs, able to understand context and to communicate effectively. The author proposes Bayesian Networks as a unifying framework to model perceptual intelligence, interpretive intelligence (user and context modeling), and narrative intelligence. Three applications have been presented to illustrate space intelligence: browsing a 3-D virtual world with natural gestures in City of News; identifying visitors' preferences and types for museum visits assisted by mobile storytelling devices; and sto(ry)chastics a real-time content selection and delivery technique which takes into account the user profile, and measurements about his behavior in the space.

The applications here described represent incremental steps towards a full implementation of space intelligence as outlined in section 3. The work carried out so far highlighted and confirmed several advantages of the proposed Bayesian modeling technique. It is:

- 1. Robust: Probabilistic modeling allows the system to achieve robustness with respect to the coarse and noisy sensor data.

- 2. Flexible: it is possible to easily test many different scenarios by simply changing the parameters of the system.

- 3. Reconfigurable: it is easy to add or remove nodes and/or edges from the network without having to "start all over again" and specify all the parameters of the network from scratch. This is a considerable and important advantage with respect to hard coded or heuristic approaches to user modeling and content selection. Only the parameters of the new nodes and the nodes corresponding to the new links need to be given. The system is extensible story-wise and sensor-wise. These two properties: flexibility and ease of model reconfiguration allow for example the system engineer, the content designer, and the exhibit curator to work together and easily and cheaply try out various solutions and possibilities until they converge to a model which satisfies all the requirements and constraints for their project. A network can also rapidly be reconfigured for other purposes.

- 4. Readable: Bayesian networks encode qualitative influences between variables in addition to the numerical parameters of the probability distribution. As such they provide an ideal form for combining prior knowledge and data. By using graphs, not only it becomes easy to encode the probability independence relations amongst variables of the network, but it is also easy to communicate and explain what the network attempts to model. Graphs are easy for humans to read, and they help focus attention, especially when a group of people with different backgrounds works together to build a new system. In this context for example, this allows the digital architect, or the engineer, to communicate on the same ground (the graph of the model) with the museum exhibit curator and therefore to be able to encapsulate the curator's domain knowledge in the network, together with the sensor data.

Future work will conduct further testing of the proposed intelligence model in a more complex space that requires the use of multiple sensors and sensor modalities to observe the behavior of people in it.

6. Bibliography

[1]. Albrecht D.W., Zukerman I., Nicholson A.E., and Bud A. "Towards a bayesian model for keyhole plan recognition in large domains." In Jameson, A.; Paris, C.; and Tasso, C., eds., *Proceedings of the Sixth International Conference on User Modeling (UM '97)*, pp. 365-376. Springer, 1997.

[2]. Azarbayejani A., Pentland, A. "Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features." In: *Proceedings of the 13th ICPR*, Vienna, Austria, 1996.

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

[4]. Brainard D.H. and Freeman W.T. "Bayesian Color Constancy." *Journal of the Optical Society of America*, A, 14(7), pp 1393-1411, July 1997.

[5]. Brand, M., Oliver, N., and Pentland, A. "Coupled hidden Markov models for complex action recognition." In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994-999, Puerto Rico , 1997.

[6]. Brooks, R.A., Coen, M., Dang, D., DeBonet, J., Kramer, J., Lozano-Perez, T., Mellor, J., Pook, P., Stauffer, C., Stein, L., Torrance, M., Wessler, M. "The Intelligent Room Project." In: *Proceedings of the Second International Cognitive Technology Conference (CT'97)*. pp. 271-279. Aizu, Japan, 1997.

[7]. Brumitt B., Meyers B., Krumm J., Kern A., and Shafer S., "EasyLiving: Technologies for Intelligent Environments", In: *Proceedings of Second International Symposium on Handheld and Ubiquitous Computing* (HUC 2000), September 2000.

[8]. Campbell, L.W., Becker, D.A., Azarbayejani, A., Bobick A., Pentland, A. "Invariant features for 3-D gesture recognition." In: *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, 1996.

[9]. Cohen M., "Design Principles for Intelligent Environments." In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI'98)*, Madison, WI. 1998.

[10]. Conati, C., A. Gertner, K. VanLehn, and M. Druzdzel. "On-Line Student Modeling for Coached Problem Solving Using Bayesian Networks." In: *Proceedings of 6th International Conference on User Modeling (UM97)*, 1997. Chia Laguna, Sardinia, Italy, 1997.

[11]. Emiliani P. L., Stephanidis C. "Universal access to ambient intelligence environments: Opportunities and challenges for people with disabilities." *IBM SYSTEMS JOURNAL*, VOL 44, NO 3, 2005

[12]. Hanssens, N., Kulkarni, A., Tuchinda, R., Horton, T. "Building Agent-Based Intelligent Workspaces." In: *Proceedings of the 3rd International Conference on Internet Computing*, pp. 675-681, 2005

[13]. Heckerman, D. “Probabilistic Similarity Networks.” Technical Report, STAN-CS-1316, Depts. of Computer Science and Medicine, Stanford University, 1990.

[14]. Howard, R.A., and Matheson, J. E. “Influence Diagrams.” In: *Applications of Decision Analysis*, volume 2, eds. R.A. Howard and J.E. Matheson, 721-762, 1981.

[15]. Jameson, A. “Numerical uncertainty management in user and student modeling: An overview of systems and issues.” In: *User Modeling and User-Adapted Interaction*, 5:193--251, 1996.

[16]. Jebara, T., and Pentland, A. “Action reaction learning: Analysis and synthesis of human behaviour.” IEEE Workshop on The Interpretation of Visual Motion at the Conference on Computer Vision and Pattern Recognition, CVPR, June 1998.

[17]. Jensen, F.V. *An Introduction to Bayesian Networks*., UCL Press, 1996.

[18]. Jensen, F.V. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

[19]. Johanson B., Fox A., Winograd T. “The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms.” *IEEE Pervasive Computing Magazine* 1(2), April-June 2002.

[20]. Jojic N., Brumitt B., Meyers B., et al. “Detection and Estimation of Pointing Gestures in Dense Disparity Maps.” In: Proceedings of *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[21]. Jordan M.I., editor. *Learning in Graphical Models*. The MIT Press, 1999.

[22]. Kidd, C. "The Aware Home: A Living Laboratory for Ubiquitous Computing Research". In: *Proceedings of the Second International Workshop on Cooperative Buildings* - CoBuild'99 October 1999.

[23]. Koller, D. and Pfeffer, A. “Probabilistic Frame-Based Systems.” In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence*; Madison, Wisconsin; July 1998.

[24]. Krumm J., Shafer S., and Wilson A., "How a Smart Environment Can Use Perception", In: *Workshop on Sensing and Perception for Ubiquitous Computing* (part of UbiComp 2001), September 2001.

[25]. Nefian A., Liang L., Pi X., Liu X. and Murphy K. “Dynamic Bayesian Networks for Audio-Visual Speech Recognition”. *EURASIP, Journal of Applied Signal Processing*, 11:1-15, 2002

[26]. Pavlovic V., Rehg J., Cham TJ, and Murphy K. “A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamic Models.” In: *Proceedings of Int'l Conf. on Computer Vision* (ICCV) 1999.

[27]. Pavlovic, V.I., Sharma, R., Huang T.S. “Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI, 19(7): 677-695, 1997.

[28]. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[29]. Pentland, A. “Smart Room, Smart Clothes.” In: *Proceedings of the Fourteenth International Conference On Pattern Recognition*, ICPR'98, Brisbane, Australia, August 16-20, 1998.

[30]. Pynadath D. V. and Wellman M. P. “Accounting for context in plan recognition, with application to traffic monitoring.” In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI, Morgan Kaufmann, San Francisco, 1995, pp. 472-481, 1995.

[31]. Rabiner, L.R. and Juang, B.H. “An introduction to hidden Markov Models.” *IEEE ASSP Magazine*, pp 4-15, January 1986.

[32]. Smyth, P. “Belief Networks, Hidden Markov Models, and Markov Random Fields: A Unifying View.” *Pattern Recognition Letters*, 1998.

[33]. Sparacino F., Wren C.R., Pentland A., Davenport G. “HyperPlex: a World of 3D Interactive Digital Movies.” In: *IJCAI'95 Workshop on Entertainment and AI/Alife*, Montreal, August 1995

[34]. Sparacino F., Pentland A., Davenport G., Hlavac M., Obelnicki M. “City of News.” *Proceedings of the: Ars Electronica Festival*, Linz, Austria, 8-13 September 1997

[35]. Sparacino F., Larson K., MacNeil R., Davenport G., Pentland A. “Technologies and methods for interactive exhibit design: from wireless object and body tracking to wearable computers.” In: *Proceedings of International Conference on Hypertext and Interactive Museums*, ICHIM 99, Washington, DC, Sept. 22-26, 1999

[36]. Sparacino F., Davenport G., and Pentland A. “Media in performance: Interactive spaces for dance, theater, circus, and museum exhibits.” *IBM Systems Journal* Vol. 39, Nos. 3 & 4, Issue Order No. G321-0139, pp. 479-510, 2000.

[37]. Sparacino F. “(Some) computer vision based interfaces for interactive art and entertainment installations.” In: *INTER_FACE Body Boundaries*, issue editor: Emanuele Quinz, Anomalie n. 2, Paris, France, Anomos, 2001.

[38]. Sparacino, F. “The Museum Wearable: real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences.” In: *Proceedings of Museums and the Web* (MW2002), April 17-20, Boston, 2002a

[39]. Sparacino F., Wren C., Azarbayejani A., Pentland A. “Browsing 3-D spaces with 3-D vision: body-driven navigation through the Internet city.” In: *Proceedings of 3DPVT: 1st International Symposium on 3D Data Processing Visualization and Transmission*, Padova, Italy, June 19-21 2002b

[40]. Sparacino F. “Sto(ry)chastics: a Bayesian Network Architecture for User Modeling and Computational Storytelling for Interactive Spaces.” In: *Proceedings of Ubicomp, The Fifth International Conference on Ubiquitous Computing*, Seattle, WA, USA, 2003.

[41]. Sparacino F. “Museum Intelligence: Using Interactive Technologies For Effective Communication And Storytelling In The Puccini Set Designer Exhibit.” In: *Proceedings of ICHIM 2004*, Berlin, Germany, August 31-September 2nd 2004.

[42]. Starner, T. and Pentland, A. “Visual Recognition of American Sign Language Using Hidden Markov Models.” In: *Proc. of International Workshop on Automatic Face and Gesture Recognition* (IWAFGR 95). Zurich, Switzerland, 1995.

[43]. Starner, T., Mann, S., Rhodes, B., Levine J., Healey, J., Kirsch, D., Picard, R., and Pentland A., “Augmented Reality through Wearable Computing.” *Presence*, Vol. 6, No. 4, pp. 386-398, August 1997.

[44]. Wren C., Basu S., Sparacino F., Pentland A. “Combining Audio and Video in Perceptive Spaces.” In: *Managing Interactions in Smart Environments* (MANSE 99), Trinity College Dublin, Ireland, December 13-14 1999

[45]. Wren C.R., Sparacino F. et al. “Perceptive Spaces for Performance and Entertainment: Untethered Interaction using Computer Vision and Audition.” *Applied Artificial Intelligence (AAI) Journal*, June 1996.

[46]. Wren, C., Azarbayejani A., Darrell, T., Pentland, A., “Pfinder: Real-Time Tracking of the Human Body.” *IEEE Trans. Pattern Analysis and Machine Intelligence.* PAMI, 19(7): 780-785, 1997.

[47]. Wu, Y. and Huang, T.S. “Human Hand Modeling, Analysis and Animation in the Context of Human Computer Interaction.” In: *IEEE Signal Processing Magazine*, Special Issue on Immersive Interactive Technology, May 2001.

[48]. Young, S.J. Woodland P.C., and Byme W.J. HTK: *Hidden Markov Model Toolkit*. V1.5. Entropic Research Laboratories Inc., 1993.