Sensory Augmented Computing: Wearing the Museum's Guide

Bernt Schiele
ETH Zurich

Tony Jebara
MIT Media Laboratory

Nuria Oliver
Microsoft Research

Abstract

A wearable computing device is much more than its desktop counterpart. It is rather like an intelligent assistant that always accompanies you and helps you solve everyday tasks. A key aspect of such devices is that they can operate autonomously and perceive the world as a human user does; without being fed manual input. They augment the user without encumbering him. A smart wearable computer sees what you see and hears what you hear to analyze, recognize and respond to the situations and people you encounter. We describe one incarnation of this smart perceptual remembrance agent, which is equipped with the ability to recognize objects in the user's visual field of view using real-time computer vision. Once an object is recognized, the system displays multimedia information that the user previously identified as being relevant to the object. The computer effectively becomes a tour guide, chiming in with augmented reality to give you reminders as you go about your routine.

1 Introduction

To date, personal computers have not lived up to their name. Most machines sit on the desk and interact with their owners for only a small fraction of the day. Smaller and faster notebook computers have made mobility less of an issue, but the same staid user paradigm persists. Wearable computing hopes to shatter this myth of how a computer should be used. A personal computer should be worn, much as eyeglasses or clothing are worn, and continuously interact with the user based on the context of the situation. With heads-up displays, unobtrusive input devices, personal wireless local area networks, and a host of other context sensing and communication tools, the wearable computer may be able to act as an intelligent assistant.

In the near future, the trend-setting professional may wear several small devices, perhaps literally built into their clothes. That way, the person may conveniently check messages, finish a presentation or browse the web while sitting on the subway or waiting in line at a bank. Such wearable devices may enhance the person's memory by providing instant access to important information anytime anywhere. Operating these devices however will be an important issue. Often today's computers require your full attention and both hands to be operated. You have to stop everything you are doing and concentrate on the device. Using speech for input and output will become more popular but may be quite annoying in many situations. Imagine for example your neighbor on a cross-Atlantic flight constantly talking and chatting with his or her devices.

Wearable devices promise to be less disruptive, and may interact with people differently from other tools. A computational device that is with you all the time can influence the sense of who you are and what you can do. Just as we have adapted to cellular phones, watches and other personal devices, wearable computers are likely to shape our personal habits around them. Starting with technophiles and migrating to the average person, culture over time will shift to incorporate them. It is too early to tell which approach to wearable design will prove popular. The devices can be built in many ways, and it will take a fashion and style battle to determine what people really want to buy.

Although their potential is vast, many of these devices suffer from a common problem: they are mostly oblivious to you and your situation. They don't know what information is relevant to you personally or when it is socially appropriate to "chime in." The goal in solving this problem is to make electronic aids that behave like a well-trained butler or an intelligent assistant. They should be aware of the user's situation and preferences, so they know what actions are appropriate and desirable - a property we call "situation awareness." They should also make relevant information available before the user asks for it and without forcing it on the user-a feature we call "anticipation and availability."

An important aspect of a wearable device is that it can perceive the world from a first-person perspective: a wearable camera can see what you see and a wearable microphone can hear what you hear in order to analyze, model and recognize things and people which are around you. A promising direction for interaction with wearable devices is therefore to make the computers more aware of the situation the user is in and to model the user's context. Sensors, such as cameras, mounted to the user's glasses, can recognize what the user is looking at and might model what the user is doing. Using sensors of various types, the device can also monitor the user's choices and build a model of his or her preferences. A person can actively train the computer by saying, "Yes, that was a good choice; show me more," or "No, never suggest country music to me." The models can also work solely by statistical means, gradually compiling information about the user's likes and dislikes, and coupling those preferences to the context.

For anticipation and availability, the wearable device can take a few key facts about the user's situation to prompt searches through a digital database or the World Wide Web. The information obtained in this manner would then be presented in an accessible, secondary display outside the user's main focus of attention.

Therefore, a key challenge for wearable computing is to model and recognize the context of the user and the situation. This contextual information is one way to achieve seamless interaction with the user. In this paper we describe a system which uses a head-mounted camera to record and analyze the visual environment of the user. In particular a computer vision program is employed to recognize objects the user is looking at. That way the system can hypothesize which part of the visual environment is interesting to the user and may display information about it when appropriate.

Several compelling attempts to provide context sensing in wearable devices have been shown. Starner [9] demonstrated a system that tracks the user's hands with a head-mounted camera to recognize American Sign Language gestures that trigger a speech synthesizer. Jebara [3] presented a wearable system that tracks a billiard table and balls to determine the best shot angles and render it on a head-mounted display. Clarkson [1] presents a system which has auditory context awareness to mediate the interface of a wearable computer. These systems are particularly effective since they bridge the interface gap between a computer and the user's environment using audio or video modalities and permit a tight autonomous coupling between computation and the user's real-world context [2] [7].

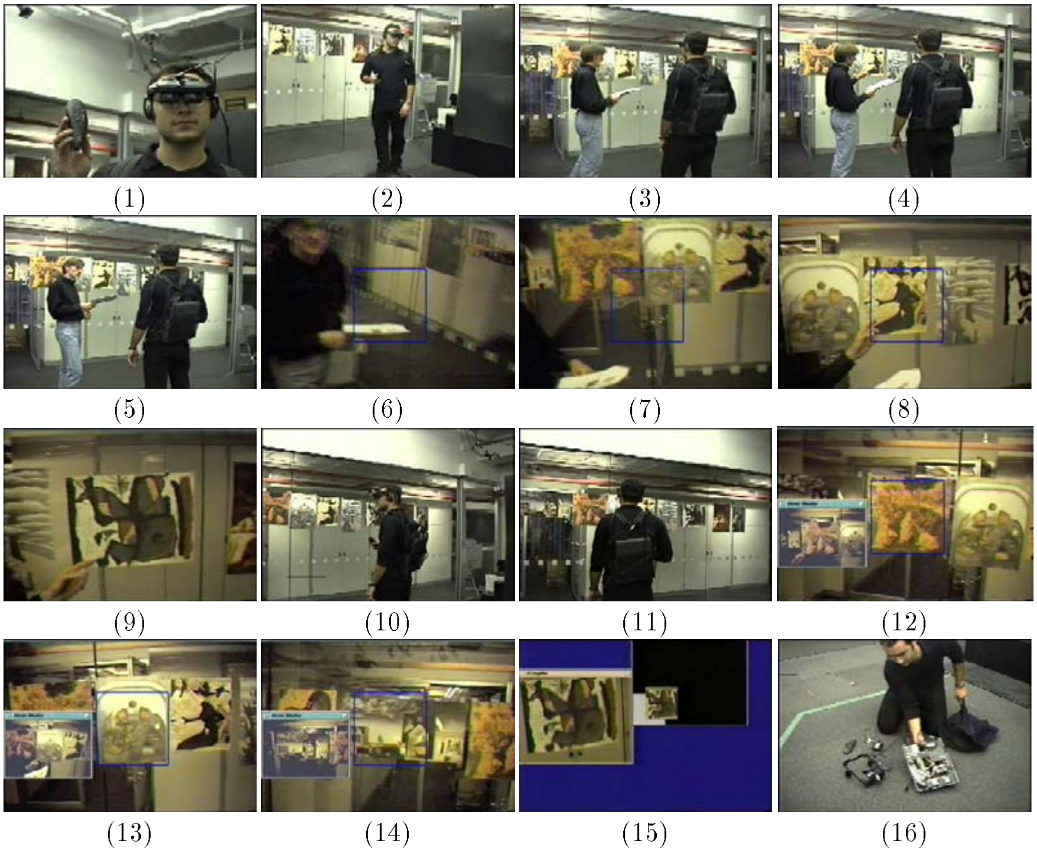
2 Wearing the Museum's Guide

Before diving into the details of our system, we will describe one compelling application for it as a personal guide for an Art Museum. Consider the following scenario: A user enters a museum-gallery with a sensory augmented wearable computer. The user walks around following a human guide which explains the works, paintings, sculptures with a narrative about each piece as well as some gestures, like pointing out some details with his hand. A museum is a rich visual environment where each piece or painting is also accompanied with many facts and details, requiring some expertise on the behalf of the tour guide. Could we replace the basic functionality of the tour guide with a perceptually smart wearable computer?

Imagine, for example, that you walk around in a museum and record video clips of a guide's explanation of the paintings as he does his tour. This is done with the help of the tiny video camera and microphone that are mounted on your wearable computer. As you record the video clips, you signal the computer to note the context you are in and associate it to the current recording. Here, the 'context' is the painting you are currently looking at or the particular corner of a room in the museum. If at any future time, the wearable's computer vision algorithm recognizes this painting or context, it will immediately replay (on your heads-up display) the video-clip that was associated with it. As you follow your guide, you slowly build an intelligent database of all the paintings. Later, when you are alone and revisit that area of the museum or painting, you are treated to a playback of the appropriate animation clip of the guide's explanation. Figure 1 depicts a real example of this process with our audio-visual remembrance agent wearable computer [1].

The wearable system is capable of this and many other scenarios and can robustly discriminate between dozens of paintings or 3d objects. If combined with further sensing, such as GPS localization, we can switch between databases (museum, office, home) to index many more objects and render many more animations. The recognized object labels could also be used to index into the database of the museum, retrieve more information about the particular painting, find similar paintings in the gallery or showing other paintings from the same painter. The system could also compute other simple statistics from the recognized objects. For example, the distribution of what objects the user looked at and how long each was gazed at. This information can be used in many ways. The system could model the preferences of the user and note the types of paintings he or she is most interested in. If we assume that the amount of time a user looks at a painting is correlated with the user's interest in it, the system can profile the interests of the user and identify, for example, other users which shared a similar profile. Alternatively, a painting that was gazed on for a long duration might trigger the system to deliver more information about it from a more detailed database. Depending on the profile, the system could then suggest other paintings in the museum. The museum could also use a

[1] Alternatively, one could have the museum's staff collect the animations and associations data for you a priori and transmit it to your wearable. This would prevent each user from customize the context of the augmentation but would not require the tour guide's help whatsoever.



(1) (2) (3) (4)

(5) (6) (7) (8)

(9) (10) (11) (12)

(13) (14) (15) (16)

Figure 1: Training the wearable computer to act as a museum guide. (1) The user with the heads-up display and a 3-button mouse to control the system. (2) The user near several paintings in a gallery. (3-5) The guide enters the scene and points to a painting while describing it. (6-7) The guide and the scene from the user's (wearable camera) field of view. (8) The guide describes a work and points to the painting while the user records the A/V. The user clicks to associate the A/V with the image of the painting at the end of that description. (9) Another work is described, recorded and associated to the painting's image. (10-14) The user walks around the space and looks at the paintings which automatically trigger a small animation on the left window. This shows the guide explaining the works appropriately with audio and his hand gesturing over the painting. (15) The object recognition engine, demonstrating robustness in recognition to changes in pose, lighting, view angle and translation. (16) The user removing the wearable system. Note, here the visor being used is a Virtual I/O prototype, we are currently using the much small MicroOptical display shown in the Hardware section.

database of user-profiles and collaborative filtering techniques to give suggestions to new visitors or to analyze the organization or effectiveness of a particular exhibit.

These additional statistics that can be derived from the user's gaze patterns have not yet been throughly investigated yet but we believe such extensions will greatly leverage the usefulness and usability of wearable computing devices. We also believe that the use of wearable sensors such as head-mounted cameras or wearable microphones combined with software to model and recognize the user's situation and context may fundamentally change human-computer interaction in general.

3 System's Overview

The building blocks of the perceptual remembrance agent are depicted in figure 2. This section describes the audio-visual association module, the object recognition algorithm used and gives a short overview of the hardware.

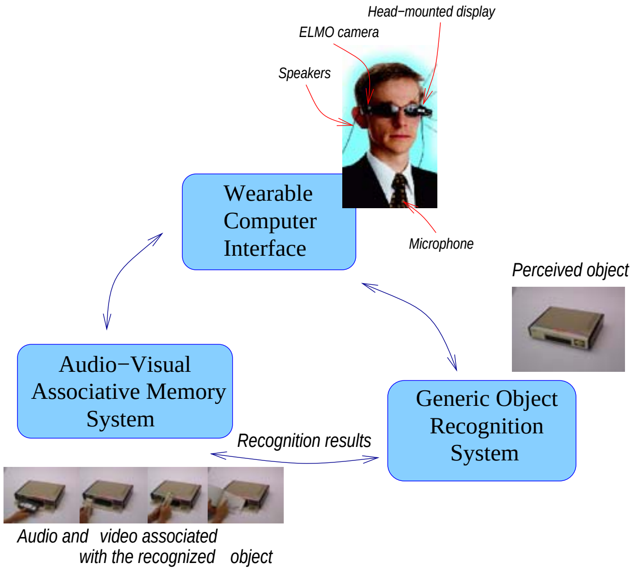

Head−mounted display
ELMO camera
Speakers
Microphone
Wearable Computer Interface
Perceived object
Audio−Visual Associative Memory System
Generic Object Recognition System
Recognition results
Audio and video associated with the recognized object

Figure 2: System's architecture

3.1 Audio-Visual Associative Memory system

The audio-visual associative memory module receives object labels along with confidence levels from the object recognition system. If the confidence is high enough, it will retrieve from memory the audio-visual information associated with the object the user is currently looking at and it will overlay this information on the real imagery that the user is perceiving.

The audio-visual recording module accumulates buffers containing audio-visual data. These circular buffers contain several seconds of compressed audio and video. Whenever the user decides to record the current interaction, the system stores the data until the user signals the recording to stop. The user moves his head mounted video camera and microphone to specifically target and *shoots* the footage desired. Thus, an audio-video clip is formed. After recording such an audio-video clip, the user selects the object that should trigger the clip's playback by directing the head-mounted video camera towards an object of interest and triggering the unit (i.e. pressing a button). The system then instructs the vision module to add the captured image*(s)* to its database of objects and associate the object's label to its most recent audio-visual clip. Additionally, the user can indicate negative interest in objects which might get misinterpreted by the vision system as trigger objects (i.e. due to their visual similarity to previously encountered trigger-objects). Thus, both positive and negative reinforcement can be performed in forming these associations. Therefore the user can actively assist the system to learn the differences between uninteresting objects and trigger objects.

The primary functionality of the perceptual remembrance agent can be projected on a simple 3 button interface (using for example a wireless 3-button mouse or a simple 3-command speech interface): a record button, an associate button and a garbage button. The record button stores the A/V sequence. The associate button merely makes a connection between the currently viewed visual object and the previously recorded sequence. The garbage button associates the current visual object with a NULL sequence indicating that it should not trigger any play back. This helps resolve errors or ambiguities in the vision system which can quickly learn when it makes an error. The association process is shown in Figure 3.

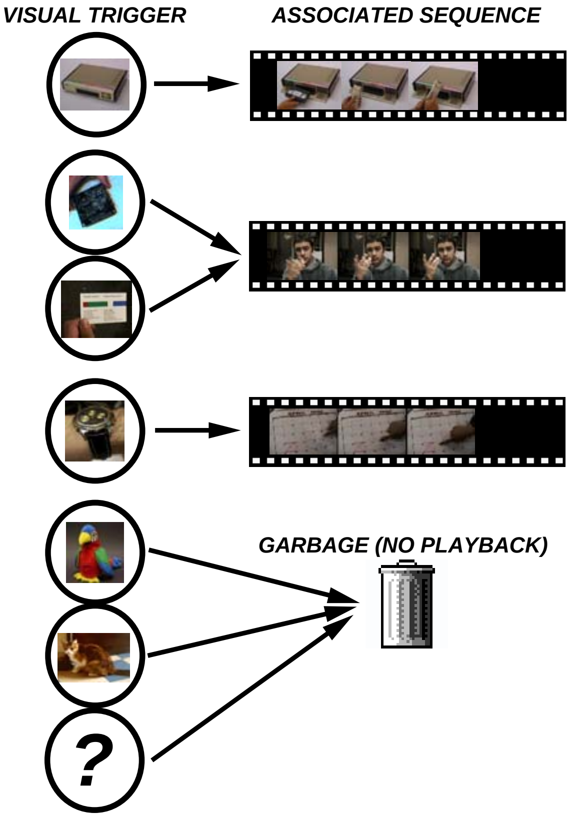


Figure 3: Associating A/V Sequences to Objects Recognized

Whenever the user is not recording or associating, the system is continuously running in a background mode trying to find objects in the field of view which have been associated to an A/V sequence. The system acts in consequence as a parallel perceptual remembrance agent that is constantly trying to recognize and explain

– by remembering associations – what the user is paying attention to. Figure 2 depicts an example of the process. Here, at an earlier time, an 'expert' demonstrated how to program a VCR. The user records the process and then associates the explanation with the image of the VCR body. Thus, whenever the user looks at the VCR he or she automatically sees an animation (overlaid on the left of his field of view) explaining how to use and program the VCR.

3.2 Object Recognition System

The input images sensed by the wearable camera are directly sent to the object recognition system. This system then tries to recognize the objects that the user is looking at. Upon recognition of some object it will send the recognition results – as object labels along with confidence levels – to the audio-visual associative memory system.

Recognizing objects is one of the most fundamental problems in computer vision and has therefore a long research history. Recognition comes at least in two flavors: "recognition of a cup, a table, a chair" or "recognition of *my* cup, *this* table or *the bedroom's* chair". The first case is typically referred to as "classification of objects" whereas the second case is called "identification of objects". Even though humans are very good at classification as well as identification of objects, the classification of objects has turned out to be an extremely difficult problem by means of machine vision. This is particularly true for unconstrained settings such as using a wearable camera in an arbitrary environment. On the other hand today it is possible to identify objects reliably even with a wearable camera – exactly what is needed for the perceptual remembrance agent. In the following we briefly describe the recognition module employed which identifies objects in real-time – an important requirement for the use in the perceptual remembrance agent.

The object recognition system used has been recently proposed by Schiele and Crowley [6]. A major result of their work is that a statistical representation based on local object descriptors provides a reliable means for the representation and recognition of object appearances. In our context this system is used to recognize previously recorded objects and to use the recognized objects as index into the audio-visual memory.

Schiele and Crowley [6] presented a technique to determine the identity of an object in a scene using multidimensional histograms [2] of vectors responses from local neighborhood operators. They showed that matching of such histograms can be used to determine the most probable object, independent of its position, scale and image-plane rotation. Furthermore they showed the robustness of the approach to view points changes.

This technique has been extended to probabilistic object recognition [6], in order to determine the probability of each object in an image only based on multidimensional receptive field histograms. Experiments showed that only a relatively small portion of the image (between 15% and 30%) is needed in order to recognize 100 objects correctly. In the following we describe briefly the technique for probabilistic object recognition. The system runs at approximately 10Hz on a Silicon Graphics machine O2 using the OpenGL extension for real-time image convolution.

Multidimensional receptive field histograms are constructed using a vector of any linear filter. Due to the generality and robustness of Gaussian derivatives, we use

[2] A histogram is a representation of a frequency distribution by means of rectangles whose widths represent class intervals and whose heights represent corresponding frequencies of occurrences or appearances of the values of the class

multidimensional vectors of Gaussian derivatives (typically the magnitude of the first derivative and the Laplace operator at two or three different scales). In order to recognize an object we are interested in the calculation of the probability of the object O_n given a certain local measurement M_k (here a multidimensional vector of Gaussian derivatives). This probability $p(O_n|M_k)$ can be calculated by the Bayes rule:

$$p(O_n|M_k) \quad = \quad \frac{p(M_k|O_n)p(O_n)}{p(M_k)}$$

with $p(O_n)$ the a priori probability of the object O_n, $p(M_k)$ the a priori probability of the filter output combination M_k, and $p(M_k|O_n)$ is the probability density function of object O_n, which differs from the multidimensional histogram of an object O_n only by a normalization factor.

Having K independent local measurements M_1, M_2, ..., M_K we can calculate the probability of each object O_n by:

$$p(O_n|M_1, \ldots, M_k) \quad = \quad \frac{\prod_k p(M_k|O_n)p(O_n)}{\prod_k p(M_k)} \tag{1}$$

M_k corresponds to a single multidimensional receptive field vector. Therefore K local measurements M_k correspond to K receptive field vectors which are typically from the same region of the image. To guarantee independence of the different local measurements we choose the minimal distance $d(M_k, M_l)$ between two measurements M_k and M_l sufficiently large (in the experiments described below we choose the minimal distance $d(M_k, M_l) \geq 2\sigma$).

In the following we assume the a priori probabilities $p(O_n)$ to be known and use $p(M_k) = \sum_i p(M_k|O_i)p(O_i)$ for the calculation of the a priori probability $p(M_k)$. Since the probabilities $p(M_k|O_n)$ are directly given by the multidimensional receptive field histograms, equation (1) shows a calculation of the probability for each object O_n based on the multidimensional receptive field histograms of the N objects. Perhaps the most tempting property of equation (1) is that we do not need correspondence. That means that the probability can be calculated for arbitrary points in the image. Furthermore the complexity is linear in the number of image points used. Therefore the recognition of objects can be done in real-time – an important requirement for the perceptual remembrance agent.

3.3 Hardware

Currently, we have demonstrated two different hardware implementations of the system. One is based on a Silicon Graphics (SGI) O2 platform which communicates with the user's camera, microphones and heads-up display via a two-way wireless radio connection. Another implementation is based on a compact Windows/PC laptop platform which is fully self-sufficient, lighter, and more affordable.

Figure 4 depicts the major peripherals that are required for the system. Output to the user is rendered via a heads-up display (HUD) which is typically the Micro-Optical clip-on 320x240 VGA device. However, earlier incarnations of the system used bulkier see-through visors such as the Sony GlassTron and the Virtual IO heads-up display. Attached to the visor is an ELMO video camera which is aligned as closely as possible with the user's line of sight [8]. Since the user has the option of viewing the world through the camera's point of view, a wide angle lens was

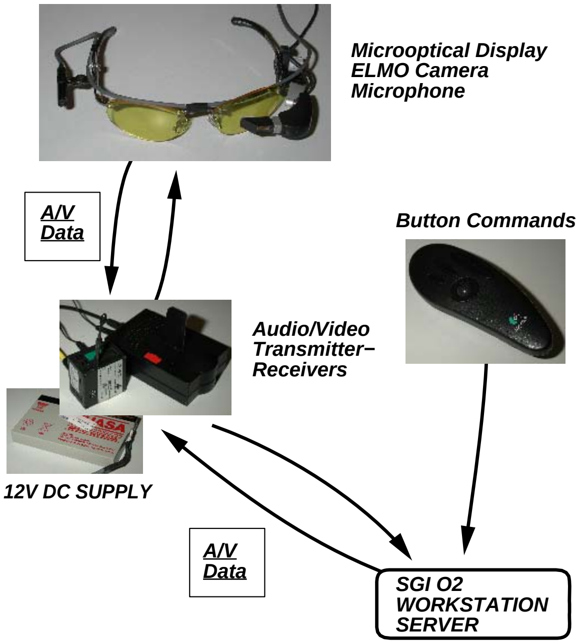


Figure 4: The Wearable Hardware System

preferred (typically 4-8mm focal lengths). In addition, a nearby microphone is also incorporated.

In the wireless system, audio/video data captured by the system is continuously broadcast using a wireless radio transmitter. The main workstation receives this information, processes the audio and video streams and sends back the processed video and audio for output onto the HUD. All communication occurs via two wireless radio transmitter-receiver pairs (different channels). This wireless transmission connects the user and the wearable system to an SGI O2 workstation where the computer vision algorithm operates. The bi-directional audio-visual connection occurs in real-time (30Hz for the video). The range of the radio frequency (RF) audio-visual link is rated at 900 feet in outdoor conditions but in most indoor and noise conditions we encountered, the user had to remain within 100 feet of the SGI base station. On board batteries permit over 8 hours of continuous operation. All components easily fit into a backpack as shown in Figure 5 weighing approximately 5lbs.

The PC-based wearable platform, being fully self-contained, allows the user an arbitrary range of movement. However, the rate of processing is slowed down considerably from the SGI due to the reduced computational power of the compact laptop. However, current sub-compact laptops are rapidly approaching the performance levels required for real-time operation and will soon rival the SGI platform.

4 Scenarios

In this section we describe some interesting scenarios and applications that naturally fall into the record-and-associate paradigm of the perceptual remembrance agent. Four coarse categories are described with some mutual overlap. Some of the concepts

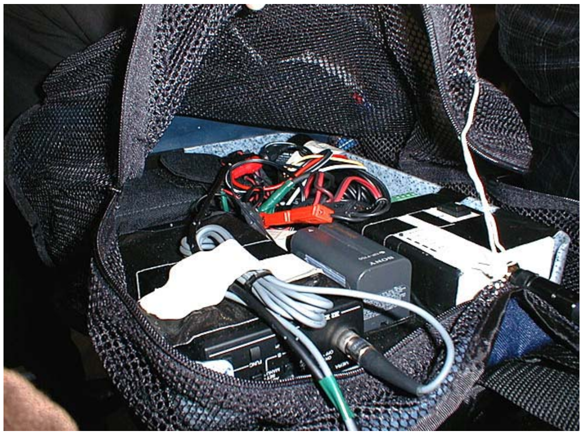


Figure 5: A Wearable Backpack

have already been implemented successfully with the system and others remain to be investigated.

4.1 Recollection of past events

One possibility is to use the system for remembering day-to-day information in an active setting. This could include daily scheduling and to-do list encoding. The user merely records his/her calendar or some notes indicating important things that need to be attended to. This recording can then be associated with the visual snapshot of the user's watch or a clock and would trigger playback whenever the user glanced at those objects. Alternatively, one could use the system to recollect a past interaction such as a communication with a business client and associate that clip with the client's business card. Whenever the user looked at the business card, the interaction would be replayed and important elements of the communication would be readily available.

4.2 Education

The system has several interesting educational applications. These introduce a subtle variation to the system's usual operation since here the recordings are performed by an expert while the learner uses the system in playback mode. For instance, the expert could be an individual with knowledge of a foreign language (i.e. French), who would use the perceptual remembrance agent to record a variety of audio pronunciations of everyday objects and to associate them with visual snapshots of the objects. Thus, a novice French learner could hear the audio playback whenever facing an object of interest and hear the corresponding French phrase. Another scenario involves having a parent posing as an expert story teller or an entertaining baby sitter and a child as the novice. The adult could read a picture book and associate each picture with the audio on the same page. The child could then enjoy hearing a story which will be synchronized to the pages of a regular every day picture book.

4.3 Online Instruction: procedural information

Consider the completion of an activity or operation which involves many sequential steps and their corresponding actions. The perceptual remembrance agent could be trained by an expert to show a novice how to perform the complex activity online and interactively. At each landmark in the activity, the expert would record the next required sub-action (which would bring the user to the following state or landmark). For instance, consider the assembly of some pre-packaged furniture. The expert associates with the fully packaged item animated instructions on how to open the box and lay out the components. Subsequently, when the vision system detects the components placed out as instructed, it would trigger the next corresponding assembly step. At each step, the system gives synchronized instructions about what to do next since the vision system is constantly tracking the evolution of the activity. In addition, if the novice performs an error and diverges from the instructions, the expert can train the system to detect this unusual state and show the user how to reverse out of this error and resume proper operation.

4.4 Augmented Perception

This category includes the variety of further sensory dimensions we may wish to incorporate to the inanimate objects we encounter. For instance, a compact disc could be associated with a small clip of the music it contains – or in a music store one could listen to the music of a CD just by looking at it; a person with poor vision could benefit by listening to an audio description of the objects in his/her field of view; in virtual advertising one could associate everyday objects with a sales pitch and for entertainment, objects could be made come to life (i.e. a plant could ask to be watered). Ultimately, the visual appearance of an object can be augmented with further audio and video of relevant messages whereas the choice and content are left to the user's imagination.

5 Summary

To date, the system has been tested by hundreds of users and has been demonstrated at several conferences and trade shows. The public events and locations where the prototype was showcased include SigGraph 1999 (USA), Darpa Image Understanding Workshop 1998 (USA), Nicograph 1998 (Japan), Heinz-Nixdorf Museum Paderborn Podium 1999 (Germany) and Orbit 2000 (Switzerland). For instance, Figure 6[3] depicts various snapshots of the wearable exhibit in Japan. Here, the system's audio-video clips play back when the user gazes at (among other things) various fashion mannequins. These clips contain explanations in Japanese, background music and footage from a fashion show exhibiting the mannequin's apparel on a real human model.

At each of these events several hundred people used the system, wandering around the exhibit area with the wearable to experience a virtual reality overlay of animations and sound on real objects and paintings in the exhibit space. Overall, most participants (from various cultures) gave positive feedback. Furthermore, the intuitive interface permitted them to quickly understand (i.e. within 1-2 minutes) the basic functionality of the device.

In a sense, the audio-visual remembrance agent is truly a personal computer since it not only records relevant information to the user, it recalls it instantly when

[3] Here, the display being used is the larger Virtual I/O display. Current versions shown in 1999 and 2000 use the small MicroOptical clip on display.

Figure 6: System Exhibit

he fixates his gaze upon an object of interest. The wearable shares the personal attention of the user and his situational context at all times. It permits a tight coupling between the natural activity of a person and the visual processing / data recollection capabilities of the machine. The computation is integrated seamlessly into everyday human activity. The availability of small wearable devices, that perceive the world, learn from the user's actions and environment and give feedback opens up a new view of computation. And with it comes a limitless array of future applications and possibilities.

References

[1] B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness via wearable computing. In *Proceedings of the Perceptual User Interfaces Workshop*, 1998.

[2] S. Feiner, B. MacIntyre, and D. Seligmann. Annotating the real world with knowledge-based graphics on a see-through head-mounted display. In *Proceedings of Graphics Interface '92*, pages 78–85, 1992.

[3] T. Jebara, C. Eyster, J. Weaver, T. Starner, and A. Pentland. Augmenting the billiards experience with probabilistic vision and wearable computers. In *Proceedings of the International Symposium on Wearable Computers (ISWC '97)*, 1997.

[4] A.P. Pentland. Wearable intelligence. *Scientific American presents: Exploring Intelligence*, 9(4), 1998.

[5] B. Rhodes and T. Starner. Remembrance agent: a continuously running automated information retrieval system. In *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology (PAAM '96)*, pages 487–495, 1996.

[6] B. Schiele and J.L. Crowley. Object recognition without correspondence using multidimensional receptive field histograms. *International Journal on Computer Vision*, 36(1):31–50, 2000.

[7] R. Sharma and J. Molineros. Role of computer vision in augmented virtual reality. In *Proceedings of SPIE - The International Society for Optical Engineering*, pages 220–231, 1995.

[8] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R.W. Picard, and A.P. Pentland. Augmented reality through wearable computing. *Presence, Special Issue on Augmented Reality*, 1997.

[9] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1372–1375, 1998.