Personal Contextual Awareness through Visual Focus

Li-Te Cheng
Lotus Research
Lotus Development Corporation
Cambridge, MA, USA, 02142
li-te_cheng@lotus.com

John Robinson
Department of Electronics
University of York
Heslington, York, UK, YO10 5DD
jar11@ohm.york.ac.uk

Abstract

This paper investigates the notion of "personal context" : an application of a wearable computer system's contextual awareness by exploiting the user's own body parts as stimuli to trigger and to act as the focus of augmentation. Our two working systems demonstrate "personal context" through the recognition of visual body cues inherent to the task via computer vision algorithms exploiting a first-person vantage point. The first system acts as a memory aid for piano playing, and the other assists novice ballroom dancing.

Keywords: humanistic intelligence, context awareness, perceptual intelligence, augmented reality, wearable computers, user interfaces, computer vision

1 Introduction

Wearable computers employing augmented and mediated reality allow their users to move freely in the environment and interact with virtual information associated with real world objects [9]. Augmentation and mediation can gather and act upon sensor readings of the wearable computer user's environment and personal activity, thus building an awareness of the user's context. Context-awareness is an important feature for wearable computer user interfaces, where computer interaction is enabled only when relevant to the current task [14].

We make a distinction between two types of wearable applications that have visual displays. The first type includes conventional PC applications as well as some context-sensitive applications. The defining character of this first type is that the display is anchored to the subject (i.e. the user or the wearer). It moves over the real world as the user moves. The second type consists of applications whose displays are at least constrained by the configuration of the real world beyond the display, if not deliberately anchored (registered) to that world. This type includes annotation applications [4], as well as other instances where overlays are spatially organised according to the environment. The display is now object-centric rather than subject-centric. We have been interested in object-centric displays because they enable the use of the whole environment as a canvas for interaction. The user's body parts are, from the point of view of display overlay, objects (i.e. potential physical anchors for interaction), yet they are under the user's intimate control and are therefore also part of the subject. So they provide a unique channel of mediation between the user and their wearable.

Sensing when it is relevant to enable interaction can be a challenge. This typically involves the analysis of the user's overall environment. Context sensing faces more complications as the wearable computer user moves around, varying the amount of available environmental computing infrastructure to enable sensing (e.g. no wireless networking coverage, going indoors with no GPS information, no "smart room" instrumented with beacons and sensors, etc). But even in the complete absence of environmental support, there remains one physical object available to mediate interaction: the user's own body. How can we exploit this to enhance a wearable computer user's experience?

We begin to answer this question in this paper, by introducing the notion of "personal context", focusing our interaction on the hands and feet. Focusing on the hands and feet can be applied to many other hand and feet oriented physical tasks: physical rehabilitation and therapy, choreography, mapping and pathfinding, sports training (e.g. martial arts, tennis, soccer, etc). Also virtual annotations and commands can be defined, moved, sized, sticked, kicked onto real world objects by hand and feet gestures. For example, framing a shot for video or a photo can triggered by a two-handed "frame" gesture, where the size and location of the framing gesture defines the parameters of the snapshot (this can also define the placement of a virtual annotation window in 3-D, like in Mann's reality window manager [9]). The wide range of suggested applications open new opportunities for mobile computing devices.

To demonstrate how personal context can enhance specific tasks on a wearable computer, we built two working systems, HANDEL and Footprint. HANDEL is a memory aid for piano playing, and Footprint assists novice ballroom dancers practicing alone.

2 Personal Context and Related Work

Personal context is the contextual awareness of the user's own body - as a stimulus and rendering surface for augmentation and mediation While a general context can be derived from the environment, a personal context can be derived from an awareness of the user's own body parts with respect to the task at hand (e.g. recognizing a physical procedure from natural hand gestures). Plus, the user's tasks often center around the active body parts, which suggest a natural focus for any virtual information presented (e.g. showing instructions near the hands in a manual task).

Thus a user-centered wearable computer system can always rely on at least one constant regardless of the current environment - the user body. Direct sensor measurements or a combination of sensors and pattern recognition can derive personal context from the user's body. Then virtual information can augment the user's first-person experience through a heads-up display, audio, projection, etc., but only enabled (or mediated) by the relevance to the user's task at hand.

Augmented reality systems [1] also overlay virtual information onto the real world, including first-person applications using head-mounted displays and environmental sensor cues to register the information onto appropriate objects to help direct a user in a task, like servicing a printer in [6] or reading an enhanced book [3]. Personal context is a niche augmented reality application, relying entirely on the user's body parts' interaction with objects and the environment to trigger virtual overlays. So a personal context approach to a printer servicing application or an augmented book would rely on the user's gaze with respect to the hands, rather than building an ultrasonic tracking infrastructure, as in [6], or using specially marked book pages, as in [3].

Personal context also relies on the user's body as a rendering surface.

This does not imply a body-stabilized interface (like a cylindrical or spherical overlay surrounding the user in [2]), but rather an object-centric interface, where the objects are really parts of the user's body, that appears world-stabilized to the user. Unlike a true world-stabilized interface as described in [2], overlays are attached to body parts with little or no attempt to assess a complete world model. So overlay graphics may be attached to the user's hands, but tracking can be done using simple 2-D techniques, with no knowledge of user's physical location, head orientation, etc. Although a complete world model is desirable, a simplified model makes available simple, fast, and perhaps robust, algorithms for tracking.

"Perceptual intelligence" [13] has a broader scope than personal context's single user's experience (e.g. "smart rooms") and can focus on environmental or user-based pattern recognition. [13] identifies new opportunities for visual contextual analysis from a first-person perspective. Specific applications include a sign-language recognizer [14] and an aid for billiards [13]. Although both cases employ body-mounted sensing to track body parts (i.e. hands gesturing or holding objects), the former lacks any augmented reality overlay and the latter also depends on some environmental awareness.

We created two systems, HANDEL and Footprint, as an initial investigation into personal context. Both systems infer the user's need for augmentation from personal context, in specific domains: piano playing and private ballroom dance practice, respectively.

3 HANDEL: Giving the User a Hand

HANDEL, a HAND based Enhancement for Learning piano music, is an example of personal context to assist learning. It uses the hands to trigger an augmented reality overlay onto the hands themselves in the context of piano playing, in essence creating a "hands-up" display.

Considerable research exists in hand-based user interfaces, as well as computer vision techniques used to locate and recognize hand and gesture, such as [13]. Hardware such as Data Gloves, magnetic trackers, and optical sensors can be used to obtain hand pose and orientation. However, in these cases, the hand acts solely as an input device. On the other hand, Krueger's work has some examples with interactive graphics merged with hands [7], and Miyasato places small displays on users' hands to ease interaction with a large screen virtual environment [10], allowing the user to "see through" the hands.

Piano teaching tools already exist in the marketplace, including self-help computer software showing keyboard and music layouts and electronic keyboards with lighted keys to guide pianists. Modern acoustic player pianos like the Disklavier allows direct playback on the keyboard from music files or from captured piano key action.

HANDEL attempts to help practicing pianists to memorize piano music. In HANDEL, the pianist equipped with a wearable computer system sits at a normal acoustic piano with no sheet music. As the pianist attempts to play a piece from memory, the pianist may look down at the hands. Focusing on the hands is the trigger for HANDEL to overlay music. Otherwise, the pianist sees nothing - no graphics clutters the practice session - and without sheet music, the pianist can concentrate on playing from memory, as if in a real recital. When the pianist looks at the right hand, only the right hand's part of the music is shown near the hand, at the current position in the piece, and similarly for the left hand. Thus, HANDEL uses the hand as an input - to trigger when or when not to overlay virtual sheet music to assist the pianist. Because the music is presented near the relevant hand, the hand also acts as context-sensitive display window for sheet music - i.e. presenting information only when needed by the pianist.

HANDEL uses a head mounted video camera to perform scene analysis, and overlays graphics on a see-through head mounted display. Thus, the pianist's hands are totally unencumbered and free to interact normally with the piano. HANDEL uses FFT phase correlation analysis [4] on consecutive video frames to determine whether the pianist's head is looking to the left or to the right. This is used to assess whether pianist is looking at the right or left hand. A look-up table skin color detection method is used to detect whether a hand is in view or not (the skin color scheme is preset with the a training set of skin color beforehand). Skin color detection is sufficient since it is assumed that the only thing the head mounted camera will see is the piano (a non-skin colored object). Figure 1 illustrates HANDEL's general system architecture and components used. On our Pentium 233 subnotebook, HANDEL runs at about 5 frames per second.

The practice session begins with the pianist loading the music score into the HANDEL program. In the current implementation, a simple, custom music score language was created to store the music in a text file. Then the pianist dons the head mounted display and sits in front of the piano. The pianist then gives a nod when starting to play the memorized music.

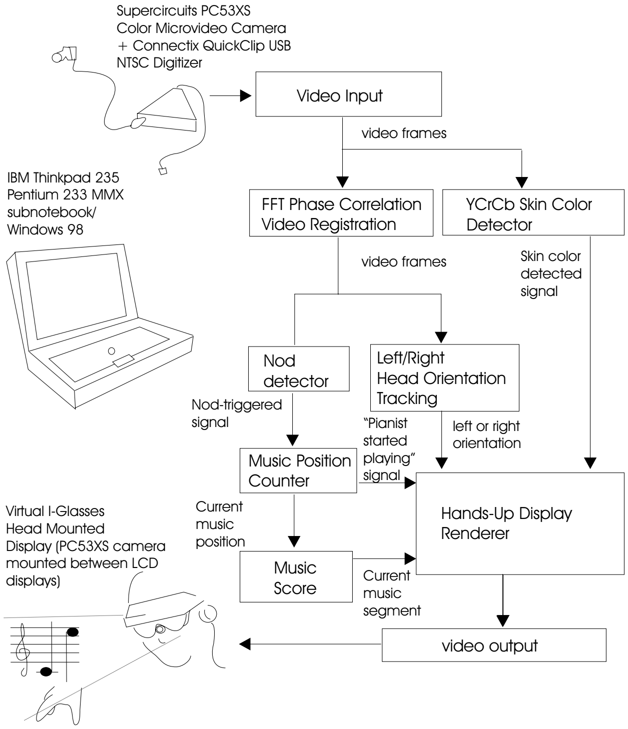

Figure 1: HANDEL system architecture and components

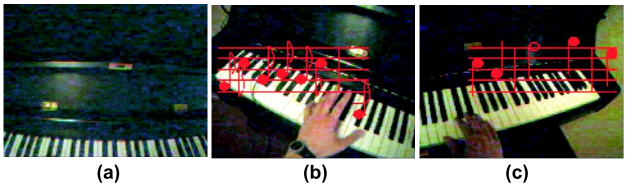

Figure 2: Views from the head mounted display - (a) nothing overlaid when no hand is in view, (b) left hand part displayed for the left hand, (c) right hand part displayed for the right hand

HANDEL uses FFT phase correlation to detect a strong vertical displacement (the nod) to begin incrementing an internal counter to keep track of the current position in the piece (in the current implementation, the counter is incremented at a predetermined rate).

While the pianist plays the piece, nothing is overlaid on the pianist's heads up display (Figure 2a) until skin color is seen by the head mounted camera. When skin is detected, the program assumes that the pianist is looking down at the hands. A specific hand is chosen based whether the pianist is looking to the left (Figure 2b) or to the right (Figure 2c). The musical score at the current position, for the given hand, is displayed on the head mounted display, and continues to update itself while the pianist is playing. The score is rendered at a fixed position on the left side of the display for the left hand, and likewise for the right hand (the score is not registered with the hand itself to avoid confusion from seeing musical notes moving with a moving hand). The virtual musical score disappears whenever the hands fall out of view (i.e. when the pianist looks up from the keys).

4 Footprint: Another Step in Personal Context

Footprint, our second personal context application, used the feet instead of the hands, as the focus for computer assistance in a ballroom dancing application.

Previous work on foot-based user interfaces fall under hardware based and computer vision based implementations. Applications for such interfaces include dance performance and choreography, motion capture for 3-D animation, and interactive entertainment.

Hardware based schemes either rely on body-mounted miniature magentic, ultrasonic, or LED devices, often monitoring the motion of the whole body. One of a few foot-specific devices include instrumented dance shoes that control music and artistic presentations [12]. Hardware systems can quickly provide great accuracy and a wealth of data, but require infrastructure or worn equipment.

Computer vision systems make use of a camera or several cameras fixed in the environment, monitoring a specific location for body motion, such as walking and running. While some systems rely on body-placed markers to aid visual detection, many analyze the scene with only an a priori model of the human body [11]. These systems are more interested in entire body motion rather than just foot motion, however. An exception is the work by [8], which derives 3-D motion data from a bicyclist's legs by analyzing specially textured shorts. Computer vision systems often free the human from wearing any special devices but need good lighting conditions and fast computers to process complex algorithms.

Footprint operates on the same minimal wearable computer system as HANDEL: a small laptop with a see-through head-mounted display with an attached video camera. A personal context is achieved by having the user's feet trigger computer interaction when they are seen. Feet detection is accomplished by analyzing the frames captured by the video camera and exploiting a priori knowledge of the wearable computer owner's feet.

Footprint's demonstration application is an aid for beginners to practice ballroom dancing steps on their own. At present, the basic waltz steps are used. Figure 3 shows Footprint's system architecture. A practice session begins with the user, equiped with the wearable computer system, starts the application and loads the system settings and dance information. An internal timer is activated, allowing Footprint to synchronize dance steps to time. The user then performs the dance to music supplied by the computer. Whenever the user needs help, s/he simply looks down at the feet. Graphics and text indicating where the feet should move next are then presented on the head-mounted display (see Figure 4). This information disappears when the user looks back up. As a consequence, looking down at the feet provides a natural means to interact with the computer. Like in HANDEL, information is only shown when needed, minimizing graphical clutter on the limited-resolution head-mounted display.

The feet detection algorithm assumes that the user is wearing dark shoes against a fairly uniform floor. Presuming that lack of edges corresponds

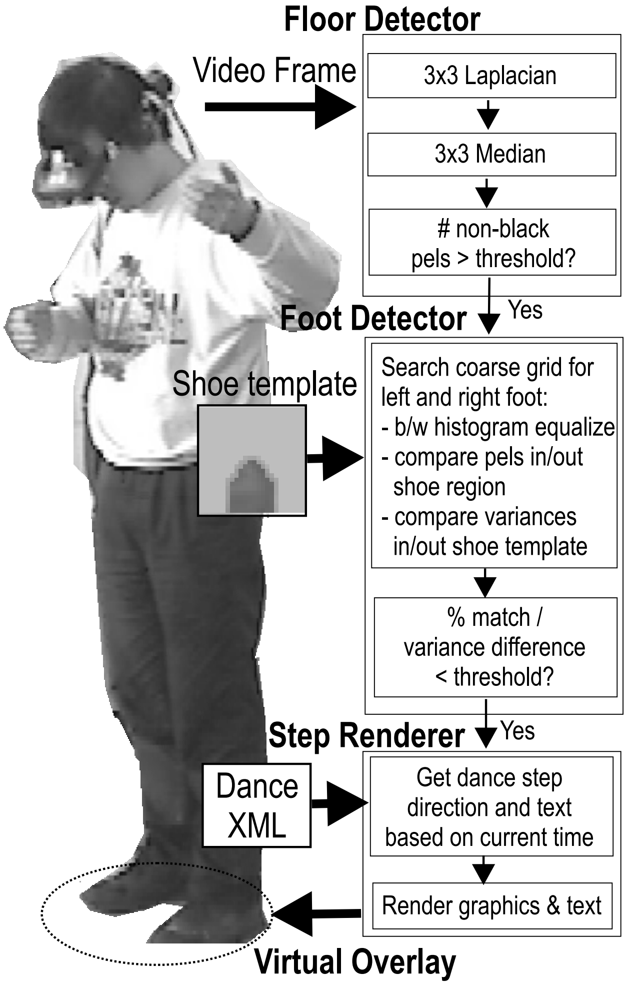

Figure 3: Footprint's system architecture

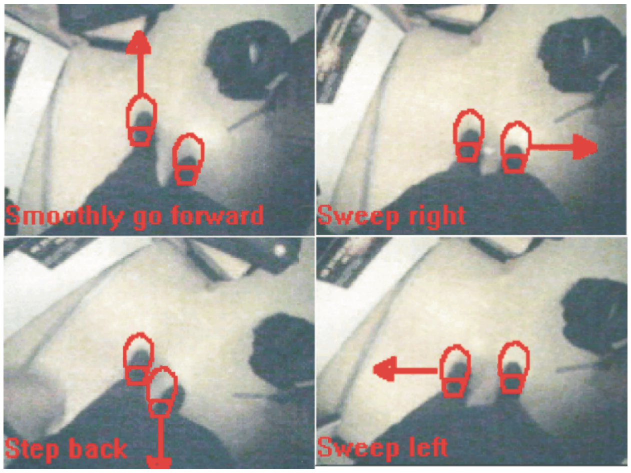

Figure 4: Dance step instructions as seen by the user's head mounted display

to uniformity, the algorithm first checks for a fairly uniform background by using a non-linear spatial activity detector. If the detector senses a cluttered background, Footprint assumes the user is not looking at a uniform floor, and will not perform any foot detection.

If the current video frame passes the "floor test", then a predefined shoe template is matched against a coarse grid on the current frame. The grid is set to the left half of the image to search for the left foot. At each grid position, the local rectangular region to be compared against the template is histogram equalized into black and white, and the number of matched pixels (with respect to the template), and local variances within the shoe and outside the shoe area are calculated. If the difference of variances between the pels within the shoe area and outside the shoe fall below a threshold (indicating the texture inside and outside the shoe are the same), or if the total difference within shoe area against the template exceeds a threshold (indicating the shoe area does not have a dark shoe), then no foot is detected. Otherwise, a measure proportional to the match against the template divided by the difference of variances is computed. The grid position with the smallest measure, and still falls under a threshold, is classified as a foot. The process then repeats itself to find the right foot, except the grid is set to the right of the discovered left foot position. Figure 5 illustrates the foot detection algorithm under different lighting and floor conditions. On subsequent steps after the first, the system searches around the last detected coordinates first

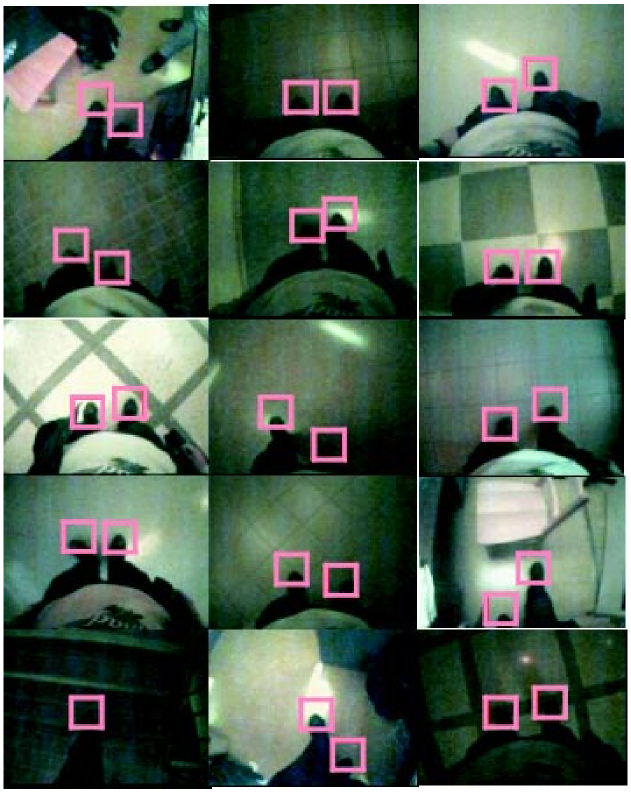

Figure 5: Footprint's foot detection under different lighting and floor conditions, as seen by the head-mounted camera. Detected feet are highlighted by rectangles

```
<dance>
  <title>Basic Waltz</title>

  <step name=”advance” duration=”quick”>
    Smoothly go forward
    <leftfoot direction=”forward”>Left first</leftfoot>
  </step>

  <step name=”right” duration=”quick”>
    Sweep right
    <rightfoot direction=”right”>Right first</rightfoot>
  </step>

  <step name=”right wait” duration=”quick”>
    Close
    <leftfoot direction=”hold”>Left arrives late</leftfoot>
  </step>

  <step name=”back” duration=”quick”>
    Step back
    <rightfoot direction=”back”>Right first</rightfoot>
  </step>

  <step name=”left” duration=”quick”>
    Sweep left
    <leftfoot direction=”left”>Left first</leftfoot>
  </step>

  <step name=”left wait” duration=”quick”>
    Close
    <rightfoot direction=”hold”>Right arrives late</rightfoot>
  </step>
</dance>
```

Figure 6: The dance markup file for the basic square-step waltz

before performing a full coarse grid search.

The dance itself is represented as an XML text file, using custom markups. As seen in Figure 6, the dance moves are clearly represented, sufficient for a ballroom dancing application. The dance steps are given in sequence, using common ballroom dance step speed denotations (“quick”, “slow”, etc). Text descriptions are provided with each movement. This new “dance markup language” is similar to SMIL, a markup language for synchronized multimedia [5]. All the parameters controlling Footprint are stored in another XML text file. The XML representation is convenient for this particular application versus a more general but complex dance notation system like Labanotation.

Footprint runs at about 4 frames per second on a Pentium 233 laptop, which includes all the image processing, video capture, and graphics rendering required by the ballroom dancing task. It detects the feet well and runs effectively with the basic waltz.

5 Discussion and Future Directions

HANDEL was tested successfully by one of the authors on an acoustic piano for a short musical piece. While it proved to be very comfortable to use, there are numerous improvements that can be made in the presentation of musical context (e.g. more informative treatment of clefs, key and time signatures, fingerings, etc). A formal user evaluation study is needed to assess HANDEL's benefit (or detriment) to memorizing piano music.

HANDEL's FFT phase approach, combined with skin detection and the assumption of a seated pianist in a front of a nearby piano, is sufficient enough for distinguishing left from right, and to detect a hand. A future improvement would be to employ affine or projective based scene analysis (such as the work specific to wearable camera systems in [9]). With a priori knowledge of a flat piano keyboard, this can form a richer interface with a pseudo 3-D world model.

Footprint can benefit from a faster computer, foot pose recognition, and further user tests to optimize the dance instruction presentation over more types of dances. Other modes for computer-assisted teaching can be explored, such as having Footprint measure the feet movements to assess a proper step. Extending the system to recognize and coordinate with a live partner would also be desirable.

Besides ballroom dancing, other foot-based personal context applications can be developed, such as for various sports and martial arts, mapping and pathfinding, physical exercise, and walking therapy for the injured and disabled, performance dance. In general, body-mounted cameras and a personalized model of the user's body running on a wearable computer promise new opportunities and new approaches for traditional vision and image processing problems.

The use of an XML based dance step file to represent content and an XML configuration file as a "style sheet" casts Footprint as a browser for a personal context wearable computer interface. Because the dance markup language is a simple description of the needed dance steps, it can be interpreted for different purposes on other platforms. For instance, another wearable computer could become create XML-based data on the fly from streaming sensor data. A 3-D capable XML desktop browser could translate the dance step file into a dancing avatar that could be incorporated into a virtual reality environment or a computer graphics movie. Online XML database engines could index and catalogue the dance step file in a repository, allowing for text-based searches for human gesture and motion. In general, context-aware applications can exploit XML as a foundation to create readable, portable, and indexable notations for human gesture, motion, and interaction with the real-world. Since gesture, motion, and interaction vary over time and depend on different conditions, context-aware notations might adapt properties and behaviours from scripting languages and temporal-based notations (such as SMIL).

Similarly the piano itself can be seen as a workspace for pianist or a composer, the dance floor can be seen as a workspace for the dancer, as the desktop is for an office worker. The piano and its sheet music, and the dance floor itself could be augmented, either by a head mounted display or an external projection system coupled with a video camera.

In conclusion, the user's attention on body parts for guidance is the basis of our demonstrations of personal context. This is a natural gesture in many tasks, thus a user can simply concentrate on the task as if the mobile computing device was not there in the first place. With only simple computer vision techniques, HANDEL and Footprint demonstrate such natural human-computer interaction in their specific application areas. And the use of XML in Footprint illustrates the potential of XML as a portable format to represent human activity in context for specialized wearable computer applications and general purpose desktop computers.

References

[1] R.T. Azuma. A survey of augmented reality. *Presence*, 6(4):355–385, August 1997.

[2] M. Billinghurst, J. Bowskill, N. Dyer, and J. Morphett. An evaluation of wearable information spaces. In *Proceedings of IEEE Virtual Reality Annual International Symposium (VRAIS 98)*, pages 20–27, Atlanta, March 14-18 1998.

[3] M. Billinghurst, H. Kato, and I. Poupyrev. The magicbook-moving seamlessly between reality and virtuality. *IEEE Computer Graphics and Applications*, 21(3):6–8, May/June 2001.

[4] L. Cheng and J. Robinson. Dealing with speed and robustness issues for video-based registration on a wearable computing platform. In *Proceed-*

ings of the Second International Symposium on Wearable Computers, pages 84–91, Pittsburgh, USA, October 19-20 1998.

[5] W3C World Wide Web Consortium. Synchronized multimedia. http://www.w3.org/AudioVideo/.

[6] S. Feiner, B. MacIntyre, and D. Seligmann. Knowledge-based augmented reality. *Communications of the ACM*, 36(7):53–62, July 1993.

[7] M.M. Krueger. Environmental technology: Making the real world virtual. *Communications of the ACM*, 36(7):36–37, July 1993.

[8] F. Lerasle, G. Rives, and M. Dhome. Tracking of human limbs by multiocular vision. *Computer Vision and Image Understanding*, 75(3):229–246, September 1999.

[9] S. Mann. Wearable intelligent signal processing. *Proceedings of the IEEE*, 86(11):2123–2151, November 1998.

[10] J. Miyasato. See-through hand. In *Proceedings of the Eigth Australian Conference on Computer Human Interaction (OzCHI 98)*, Adelaide, Australia, November 29-December 4 1998.

[11] J. Ohya, J. Kurumisawa, and R. Nakatsu. Virtual metamorphosis. *IEEE Multimedia*, 6(2):29–39, April/June 1999.

[12] J. Paradiso, K. Hsiao, A. Benbasat, and Z. Teegarden. Design and implementation of expressive footwear. *IBM Systems Journal*, 39(3 and 4):511–529, 2000.

[13] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–118, January 2000.

[14] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *Proceedings of the Second International Symposium on Wearable Computers*, pages 50–57, Pittsburgh, USA, October 19-20 1998.