MVA2002 in Nara, Japan, pp.84-89 (2002)

A Functionally-Distributed Hand Tracking Method for Wearable Visual Interfaces and Its Applications

Takeshi Kurata† Takekazu Kato† Masakatsu Kourogi†
Jung Keechul‡ Ken Endo§
†Intelligent Systems Institute
National Institute of Advanced Industrial Science and Technology (AIST), JAPAN
‡Michigan State University, USA §Media Drive Corporation, Japan
E-mail: kurata@ieee.org

Abstract

This paper describes a Functionally-Distributed (FD) hand tracking method for hand-gesture-based wearable visual interfaces. The method is an extension of the Distributed Monte Carlo (DMC) tracking method which we have developed. The method provides coarse but rapid hand tracking results with the lowest possible number of samples on the wearable side, and can reduce latency which causes a decline in usability and performance of gesture-based interfaces. The method also provides the adaptive tracking mechanism by using the sufficient number of samples and the hand-color modeling on the infrastructure side. This paper also describes three promising applications of the hand-gesture-based wearable visual interfaces implemented on our wearable systems.

1 Introduction

Computer vision is a key technology for providing post-WIMP interfaces such as AR interfaces and PUIs. In recent years, a significant number of attempts have been made to develop hand-gesture (HG) based interfaces especially for mobile or wearable systems [7, 10, 11, 12, 13, 14, 15, 16, 18]. However, many of them are often sensitive to changes in lighting conditions and background, or computationally intensive for stand-alone wearable computers whose computational resources and battery power are often limited by their size necessary for ensuring the wearability.

We previously developed a client/server type of vision-based wearable system to compensate for lack of the computational power of wearable systems [8, 9, 10]. In the systems, the wearable client processes only I/O tasks such as capturing/decoding/encoding images and displaying information, and the server processes all computer-vision tasks with image data transmitted through a wireless LAN. However, although such a server can handle many intensive tasks at high throughput, it is quite difficult to respond to the wearable client via a wireless network with minimum latency. In addition, such systems can easily stall when the wireless connection is unstable due to roaming, interference, and noise.

In this paper, we propose a Functionally-Distributed (FD) hand tracking method for *Wea*rable *V*isual *I*nterfaces (Weavy) [1]. The method provides coarse but rapid hand tracking results based on the ConDensation algorithm [2] with the lowest possible number of samples on the wearable side. As a result, it can reduce latency which causes a decline in usability and performance of gesture-based interfaces. The method also provides the adaptive tracking mechanism by using the sufficient number of samples and the hand-color modeling on the infrastructure side. The adaptive tracking mechanism is considerably intensive for the wearable-side modules, but it needs not to be processed in real time. Therefore, we can design this task as one of the XML web services on the infrastructure-side modules, and color models prefetched on the wearable-side modules are used until updated color models are received. This paper also describes three promising applications of the HG-based Weavy implemented on our wearable systems: a virtual universal remote control, a secure password input, and a real world OCR. These application tasks are also implemented on the infrastructure-side modules as the XML Web services.

2. Functionally-Distributed Hand Tracking

Figure 1 is the diagram of the Functionally-Distributed (FD) hand tracking method which we propose in this paper. The method consists of tracking tasks based on

the Distributed Monte Carlo (DMC) tracking method [5] and an adaptive color modeling task based on the color histogram approximation by means of a Gaussian mixture model (GMM) [10, 18]. Each task of the FD tracking method is assigned to wearable-side modules which are the modules worn by the wearer, or infrastructure-side modules which are all other modules.

In [5, 6], the DMC tracking method was used to track the target person by controlling a wearable active camera (WAC) with minimum latency and also was used to obtain accurate position and shape of the face. In this paper, we apply the method to HG-based Weavy, since the latency of response should be cut down to prevent a decline of usability and the adaptation to environmental changes should be satisfied for such interfaces.

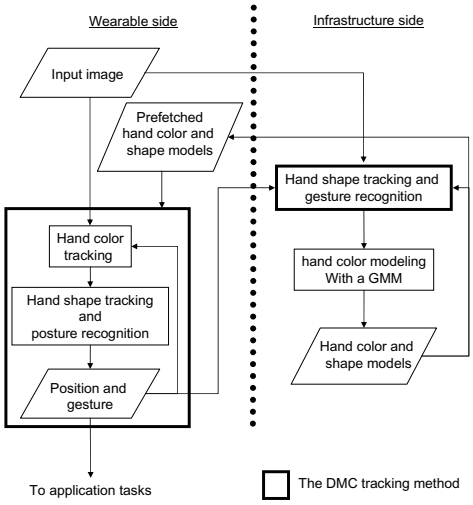


Figure 1: Diagram of the FD hand tracking method.

2.1 The DMC Tracking Method

The DMC tracking method [5] is an extension of the ConDensation algorithm [2] for distributed architectures. The iterative process of the ConDensation algorithm consists of the following three steps:

1. Select each sample $\boldsymbol{s'}_t^{(n)} (= \boldsymbol{s}_{t-1}^{(j)})$ with the weight $\pi_{t-1}^{(j)}$ $(n, j = 1, .., N)$.
2. Predict by sampling from the dynamical model $p(\boldsymbol{X}_t | \boldsymbol{X}_{t-1} = \boldsymbol{s'}_t^{(n)})$ to generate each $\boldsymbol{s}_t^{(n)}$.
3. Weight each $\boldsymbol{s}_t^{(n)}$ by means of observed features $\boldsymbol{Z}_t$: $\pi_t^{(n)} = p(\boldsymbol{Z}_t | \boldsymbol{X}_t = \boldsymbol{s}_t^{(n)})$.

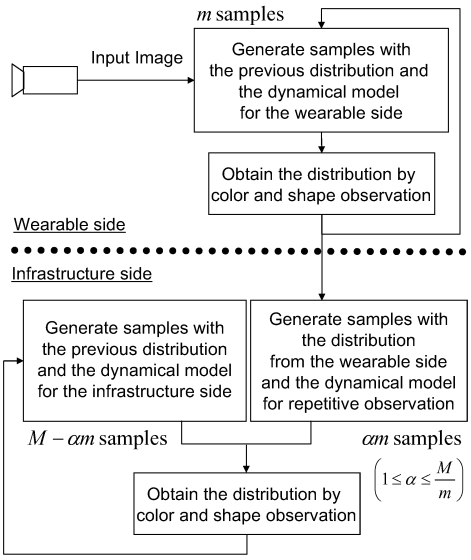


Figure 2: Diagram of the DMC tracking method.

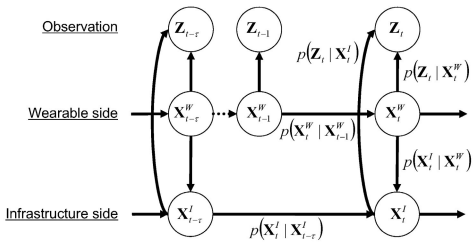


Figure 3: State transition diagram of the DMC tracking method.

.

Since the ConDensation algorithm or sequential Monte Carlo tracking methods have plural hypotheses represented with a discrete probability density $\{\boldsymbol{s}_t^{(n)}, \pi_t^{(n)}\}$, they track the target object robustly even if the object is in clutter. However, we need a large number of samples to fully bring out the performance. To solve this problem, the DMC tracking method uses different numbers of samples and different dynamical models on the wearable side and infrastructure side respectively so as to obtain the tracking result with the minimum latency on the wearable side and to accurately estimate the state of the object on the infrastructure side.

The boxes with bold lines in the diagram of the FD hand tracking method (Figure 1) correspond to the DMC tracking method, and Figure 2 explains the detail. The iterative process on the infrastructure

side generates new samples $\boldsymbol{s}_t^{I(n)}$ from both the new but sparse sample-set received from the wearable side $\{\boldsymbol{s}_t^{W(n)}, \pi_t^{W(n)}\}$ and the old but dense sample-set on the infrastructure side $\{\boldsymbol{s}_{t-\tau}^{I(n)}, \pi_{t-\tau}^{I(n)}\}$. A sample mixture parameter α in Figure 2 controls the selection ratio of $\boldsymbol{s}_t^{W(j)}$ to $\boldsymbol{s}_{t-\tau}^{I(j)}$ according to the wireless network condition.

Figure 3 shows the state transition diagram of the DMC tracking method where $p(\boldsymbol{Z}_t|\boldsymbol{X}_t^W)$ and $p(\boldsymbol{Z}_t|\boldsymbol{X}_t^I)$ indicate respectively the likelihood of observation of the state of the object $\boldsymbol{X}_t^W$ generated on the wearable side and the likelihood of the state $\boldsymbol{X}_t^I$ generated on the infrastructure side. The dynamical models $p(\boldsymbol{X}_t^W|\boldsymbol{X}_{t-1}^W)$ and $p(\boldsymbol{X}_t^I|\boldsymbol{X}_{t-\tau}^I)$ are used respectively for the wearable side and infrastructure side. $p(\boldsymbol{X}_t^I|\boldsymbol{X}_t^W)$ is also a dynamical model. However, unlike with the above two dynamical models, this model is used to generate new samples from the results obtained at the same frame (time t), so we call $p(\boldsymbol{X}_t^I|\boldsymbol{X}_t^W)$ the dynamical model for repetitive observation.

2.2 Shape Representation and Observation

We use a simple 2-D contour model for hand shape representation as shown in Figure 4 and employ the following seven parameters to describe the state:

$\boldsymbol{X}_t = (\theta_t, t_{x,t}, t_{x,t}, s_t, \phi_{x,t}, \phi_{y,t}, \gamma_t)$,

where θ_t is the rotation angle, $(t_{x,t}, t_{x,t})$ is the center position of the hand shape, $s_{x,t}$ is the scaling parameter, and $(\phi_{x,t}, \phi_{y,t})$ is the shear parameter. The posture parameter $\gamma_t (0 \leq \gamma_t \leq 1)$ controls the hand shape from the pointing posture (Figure 4 (a)) to the clicking posture (Figure 4 (b)) the same way as image morphing. By using these postures, the user can interact with wearable appliances as shown in Figure 5.

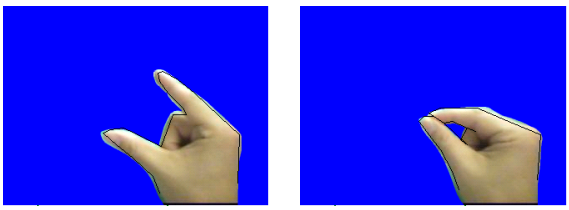
The dimensionality of the shape space is much higher compared with the number of samples that can be used on the wearable side. To follow the large motion of the hand, we design the dynamical model $p(\boldsymbol{X}_t^W|\boldsymbol{X}_{t-1}^W)$ so that each sample is distributed sparsely and over a wide range except the shear parameter. On the other hand, $p(\boldsymbol{X}_t^I|\boldsymbol{X}_{t-\tau}^I)$ is designed to estimate all parameters of X_t accurately.

Let $\boldsymbol{e}_1$ be the first eigenvector of the covariance matrix $\boldsymbol{G}$ of the image gradient (g_x, g_y) around an observation point $P_c (c = 1, .., C)$ on the contour. In our implementation, the likelihood of each sample for shape observation is defined as

$$p(\boldsymbol{Z}_t|\boldsymbol{X}_t = \boldsymbol{s}_t^{(n)}) = \prod_{c=1}^{C} p(\boldsymbol{z}_t|\boldsymbol{X}_t) \qquad (1)$$

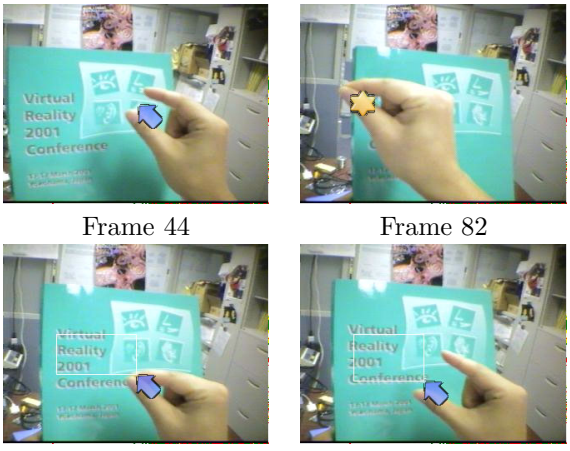
$$p(\boldsymbol{z}_t|\boldsymbol{X}_t) \propto \begin{cases} q + |\boldsymbol{e}_1^T \boldsymbol{n}_c| & \text{if } \lambda_1 > T_\lambda \\ q & \text{otherwise} \end{cases}$$

where λ_1 is the first eigenvalue of $\boldsymbol{G}$, T_λ is a threshold of λ_1, $\boldsymbol{n}_c$ is the normal vector of P_c, and q is a constant to survive a transitory failure of observations due to occlusion of the tracked object.



(a) Pointing [$\gamma = 0$] (b) Clicking [$\gamma = 1$]

Figure 4: A simple hand shape model. This model has 21 observation points on the contour.



Frame 44 Frame 82

Frame 121 Frame 129

Figure 5: Selecting a rectangle by dragging.

2.3 Color Representation and Observation

As described in [5], we use color observation before shape observation based on importance sampling [3, 17] not only to reduce samples in the background but also to reduce the computational cost of shape observation. We use a Gaussian distribution as a parametric color representation. The model is estimated on the infrastructure side as described in 2.4 and uploaded to the wearable side when wireless connection is available, so the prefetched color model is used for color observation. For color observation, each Gaussian distribution of samples is obtained inside each hand-shape contour and the similarity of the two distributions is evaluated using the Mahalanobis distance.

2.4 Color Modeling

Instead of using predefined hand-color models, we dynamically construct hand- and background-color models based on the hand-color-segmentation method proposed in [10]. The method uses a GMM to approximate the color histogram of each input image. The GMM is estimated by the restricted Expectation-Maximization (EM) algorithm in which the standard EM algorithm was modified to make the first Gaussian distribution an approximation of the hand-color distribution [18]. Not only to obtain the estimated mean of hand color necessary for the restricted EM algorithm that estimates the GMM but to classify hand pixels based on the Bayes decision theory, we need a spatial probability distribution of hand pixels. In this study, we use the hand shape estimated by the DMC tracking method as the distribution. This hand-color modeling task is assigned on the infrastructure side, and the first Gaussian distribution of the GMM is send to the wearable side as described the above.

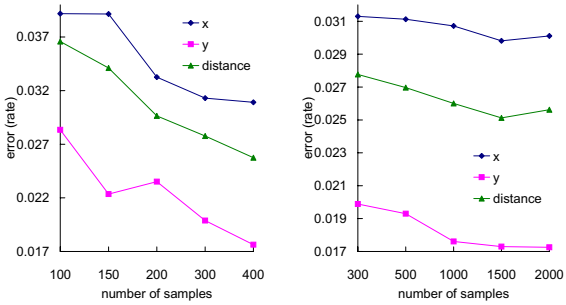
3. Experiments

We evaluated the accuracy of hand tracking using an image sequence taken with a head-worn camera. The sequence has 227 frames which includes pointing and clicking postures. In this experiment, the estimated hand state was obtained using the weighted mean of samples $\{\boldsymbol{s}_t^{(n)}, \pi_t^{(n)}\}$.

Figure 6 shows the average pointing errors for varied numbers of samples and for the wearable and infrastructure sides. The results consist of the distance along the x axis, the distance along the y axis, and the Euclidean distance between the estimated position of the thumb's tip and the ground-truth position measured manually. These results are normalized so that the width and the height of input image are 1.0 respectively. It is self-evidence that the pointing accuracy was improved as the number of samples increases.

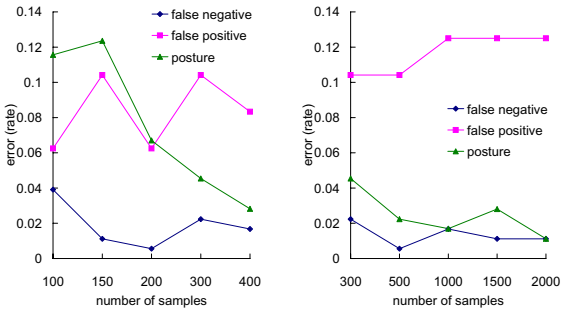
Figure 7 graphs posture classification error and also graphs false positive and false negative in estimating the existence of the hand. We classified each hand shape as pointing or clicking by simply using a threshold of γ_t and we estimated the existence of the hand by using the likelihood of color and shape observation (Figure 8). The performance of posture recognition was improved as the number of samples increases, although the false positive was not. However, since many of the false positive errors occur just after disappearing the hand, we can reduce such errors by using some simple sequential filters.

Considering a balance of the computational resource and the tracking accuracy, we currently set the number of samples to 300 for the wearable side (CPU: Crusoe 867MHz) and to 1000 for the infrastructure side (CPU: Pentium IV-M 2.2GHz). Figure 9 is an example of the trajectory of thumb's tip on the wearable side (N=300) and Figure 5 is example output based on results of this experiment.



(a) Wearable (b) Infrastructure

Figure 6: The average pointing errors for varied numbers of samples and for the wearable and infrastructure sides.



(a) Wearable (b) Infrastructure

Figure 7: Posture classification error and false positive and false negative in determining the existence of the hand.

4. Applications

We have developed several types of wearable systems provided with Weavy, called *Wyvern* [1]. Figure 10 indicates the distributed software architecture of the *Wyvern* including the HG-based interface. In this section, we will introduce three applications of the HG-based Weavy implemented on the *Wyvern* (Figure 12).

4.1. Virtual Universal Remote Control

The personal positioning method which we developed [8] can obtain the wearer's position and direction by using image registration and sensor-data fusion techniques, and display video frames overlaid with 2-D

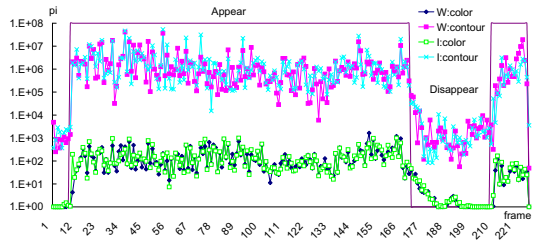


Figure 8: Likelihood at each frame. Bold line shows whether the hand exist in each image or not.

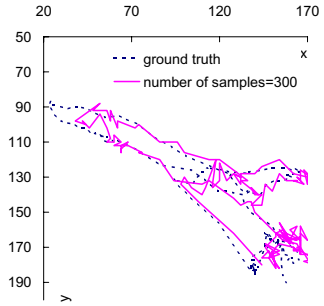


Figure 9: Trajectory of the pointer from frame 9 to 170.

annotations related to the wearer's view. Using this method, for instance, the system can show the wearer the virtual control panel of some appliance which is in its field of view, so that the wearer can operate the appliances with its own hand as shown in Figure 12 (a).

4.2. Secure Password Input

Figure 12 (b) shows some example output of the secure-password-input application with the HG-based Weavy. Since the position and the shape of each key in soft keyboard are randomly changed at every input, it is very difficult for anybody to steel the password, even if it observes the wearer's hand motion very carefully.

4.3. Real-World OCR

The scene text detection method [4] provides several candidate regions to process an OCR task. Just selecting one of those candidates with the wearer's hand, text information in the real world is acquired and can be used for various purposes such as translation, navigation service based on signboards, web search, and so on (Figure 11). Figure 12 (c) is some example output of the Real World OCR (RWOCR) and three rectangles in Figure 12 (d) show the text regions which were

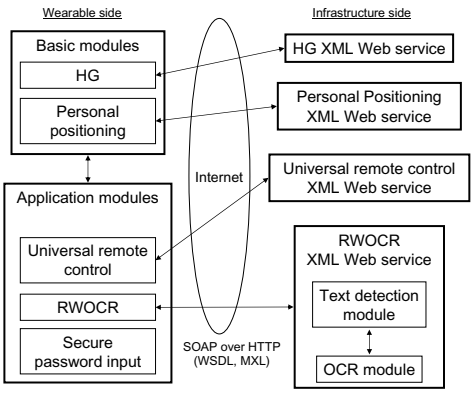


Figure 10: Distributed software architecture of the Wyvern system.

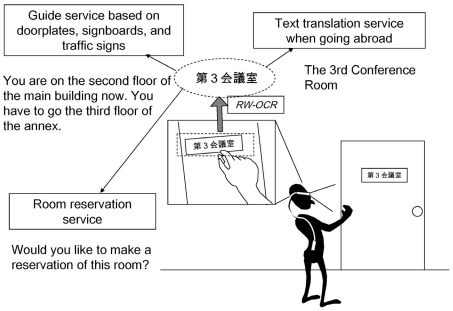


Figure 11: Overview of RWOCR.

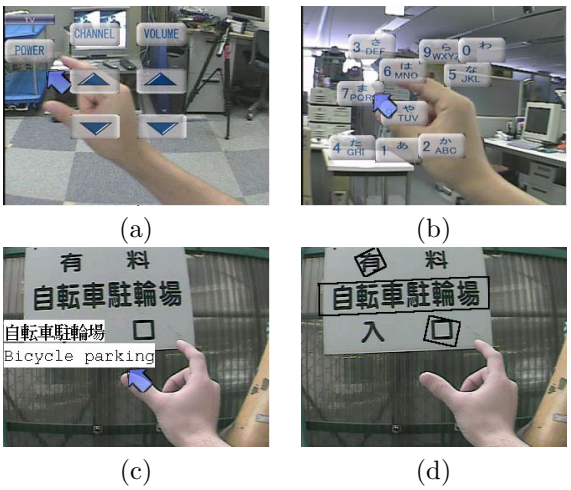
automatically detected.

5. Conclusions

We proposed the FD hand tracking method based on the DMC tracking method and the color modeling with the GMM. The method not only provides coarse but rapid hand tracking results with the lowest possible number of samples on wearable side but provides the adaptive tracking mechanism with the sufficient number of samples and the hand-color modeling on the infrastructure side. With this distributed framework, the wearable side is capable of continuing to track the hand by itself even when it is unable to communicate with the infrastructure side. Furthermore, more accurate results and the learning data are obtained when it is able to communicate with the infrastructure side.

Acknowledgements

The authors would like to thank Takashi Okuma, Hironobu Takahashi, and Katsuhiko Sakaue for their constructive discussions and help. This work is supported



(a) (b)

(c) (d)

Figure 12: Applications of the HG-based Weavy. (a) Universal remote control, (b) Secure Password Input, (c) RWOCR. (d) Automatically detected text regions.

in part by Special Coordination Funds for Promoting Science and Technology of MEXT of the Japanese Government.

References

[1] *Weavy: Wearable Visual Interfaces*, http://www.is.aist.go.jp/weavy/.

[2] M. Isard and A. Blake. Condensation – condentional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.

[3] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. 5th European Conference on Computer Vision (ECCV98)*, pages 893–908, 1998.

[4] K. Jung, K. I. Kim, T. Kurata, M. Kourogi, and J. Han. Text scanner with text detection technology on image sequences. In *Proc. 16th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 473–476, 2002.

[5] T. Kato, T. Kurata, and K. Sakaue. VizWear-Active: distributed monte carlo face tracking for wearable active camera. In *Proc. 16th International Conference on Pattern Recognition (ICPR)*, volume 1, pages 395–400, 2002.

[6] T. Kato, T. Kurata, and K. Sakaue. VizWear-Active: towards a fanctionally-distributed architecture for real-time visual tracking and context-aware UI. In *Proc. 6th Int'l Symp. on Wearable Computers (ISWC2002)*, pages 162–163, 2002.

[7] T. Keaton, S. M. Dominguez, and A. H. Sayed. Snap & tell: A vision-based wearable system to support 'web-on-the-world' applications. In *Proc. 6th Australasian Conference on Digital Image Computing Techniques and Application (DICTA2002)*, pages 92–97, 2002.

[8] M. Kourogi, T. Kurata, and K. Sakaue. A panorama-based method of personal positioning and orientation and its rea-time applications for wearable computers. In *Proc. 5th Int'l Symp. on Wearable Computers (ISWC2001)*, pages 107–114, 2001.

[9] T. Kurata, T. Okuma, M. Kourogi, T. Kato, and K. Sakaue. VizWear: Toward human-centered interaction through wearable vision and visualization. In *Proc. 2nd IEEE Pacific-Rim Conf. on Multimedia (PCM2001)*, pages 40–47, 2001.

[10] T. Kurata, T. Okuma, M. Kourogi, and K. Sakaue. The Hand Mouse: gmm hand color classification and mean shift tracking. In *Proc. 2nd Int'l Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS2001) in conjunction with ICCV2001*, pages 119–124, 2001.

[11] S. Mann. Wearable computing: A first step toward personal imaging. *Computer*, 30(2):25–32, 1997.

[12] H. Sasaki, T. Kuroda, Y. Manabe, and K. Chihara. HIT-Ware : A menu system superimposing on a human hand for wearable computers. In *Proc. 9th Int'l Conf. on Artificial Reality and Telexistence (ICAT'99)*, pages 146–153, 1999.

[13] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Proc. 4nd Int'l Symp. on Wearable Computers (ISWC2000)*, pages 87–94, 2000.

[14] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, W. R. Picard, and A. Pentland. Augmented reality through wearable computing. Technical Report 397, M.I.T Media Lab. Perceptual Computing Section, 1997.

[15] T. Starner, J. Weaver, and A. Pentland. A wearable computer based american sign language recognizer. In *Proc. 1st Int'l Symp. on Wearable Computers (ISWC'97)*, pages 130–137, 1997.

[16] A. Vardy, J. Robinson, and L.-T. Cheng. The wristcam as input device. In *Proc. 3rd Int'l Symp. on Wearable Computers (ISWC'99)*, pages 199–202, 1999.

[17] Y. Wu and T. S. Huang. A co-inference approach to robust visual tracking. In *Proc. The 8th IEEE Int'l Conf. on Computer Vision (ICCV2001)*, volume 2, pages 26–33, 2001.

[18] X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *Proc. 4th Int'l Conf. on Automatic Face and Gesture Recognition (FG2000)*, pages 446–453, 2000.