# Robust Tracking of People by a Mobile Robotic Agent

Rawesak Tanawongsuwan, Alexander Stoytchev, Irfan Essa

College of Computing, GVU Center,
Georgia Institute of Technology
Atlanta, GA 30332-0280 USA
{tee|saho|irfan}@cc.gatech.edu

February 25, 1999

**Abstract**

We present methods for tracking people in dynamic and changing environments from camera mounted on a mobile robot. We describe processes to extract color, motion, and depth information from video and we present methods to merge these processes to allow for reliable tracking of people. We discuss how this merging of different measurements can aid in instances where there is motion in the scene due to large movements by people, camera movements, lighting variations, even in the presence of skin-like colors in the scene. We also apply the results from our tracking system for gesture recognition in the context of human-robot interaction.

**Keywords:** People tracking, Gesture Recognition, Vision-based user interfaces, Robot vision.

## 1 Introduction

One of the goals of building intelligent and interactive machines is to make them aware of the user's presence. In this paper we present various computer vision methodologies for building a system capable of determining the presence of humans in a scene, tracking their locations, and recognizing their gestures in complex and dynamic environments. We present methods for tracking users that work with a moving camera, changing backgrounds, varying lighting conditions. These methods rely on techniques for color tracking, motion and depth estimation, and algorithms for combining these techniques for robust tracking of users.

Our main motivation for tracking people in dynamic and changing environments is our interest in building mobile robots that can track people and recognize their gestures and activities, and react accordingly. Unfortunately, most of the scenarios in which we would like these machines to operate in are very dynamic and unconstrained. In this paper, we present methods to combine color, motion, and depth cues for robust tracking of users under conditions of moving cameras, changing backgrounds, varying lighting conditions. We describe the methodologies to address these problems and obtain better tracking. The results of the tracker are applied for gesture recognition for human-robot interaction.

**Related Work.** There has been much work in the area of people tracking using computer vision techniques. Several real-time systems have been presented that work reliably under a few established conditions that do not

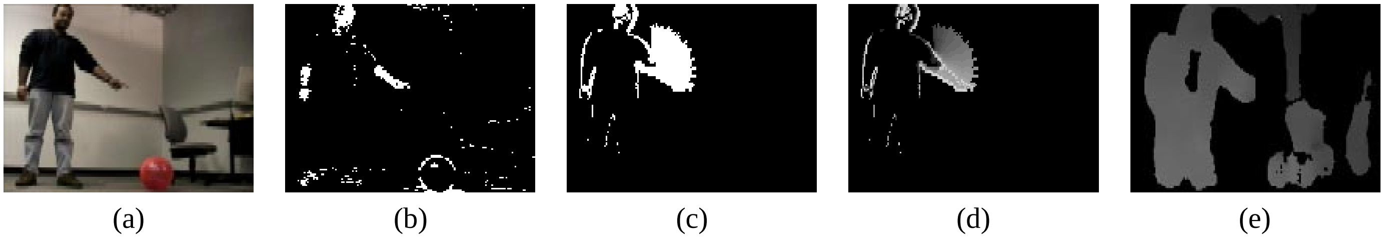


Figure 1: *Tracking of a person from a camera mounted on a robot. The person performs a pointing gesture commanding the robot to fetch a ball. (a) Original image. (b) Segmented skin color image. (c) Binary motion energy image. (d) Motion history image, (shows the direction of the motion), and (e) stereo depth image.*

change over time. Color-based tracking with a static background is the basis of the widely used people-tracking system, Pfinder [20]. Several other systems have also concentrated on similar color metrics for tracking faces in real-time [3, 7, 22].

The increase in computational power has made it possible to further extend the tracking methods to allow for multiple cameras [1] and combination of other visual cues like motion and stereo [6]. This has resulted in a variety of real-time systems that have been used by thousands of people [5, 2, 10, 14] at various exhibitions.

There has also been much progress in locating people in a scene using methods to look for faces [12, 15, 18]. These methods, though much more reliable than the color-tracking methods, are computationally quite expensive making them somewhat infeasible for real-time systems. These methods require higher resolution images than the color-tracking methods. Some attempts have been made to incorporate these methods with the real-time systems to aid in initialization and registration [5].

A major issue with most of the real-time systems is that each method works under specific conditions. For example, background subtraction and motion tracking techniques will work fine with low-resolution images, however, will fail to extract a good resolution face image for further processing (*e.g.,* face and expression recognition). To address these issues, Crowley and Berard [3] present an architecture for real-time systems where a supervisory controller selects and activates various visual processes in order to ascertain a degree of confidence concerning observed data. Similar issues are addressed by building foveating systems as presented by [11, 4, 17]. These systems use an active camera to focus on the location of the user after an initial process locates the user. In this paper, we present various methods for tracking people and propose methods of combining different processes that operate at different levels. It is these combinations of processes that allow for reliable tracking under dynamically changing conditions.

In addition to the above mentioned methods for tracking people, there has been some recent research on tracking people using a camera mounted on a robot. The processing required to track people from a robot is challenging as the conditions are not fixed and require adaptive methods to deal with the changes in the scene [19, 9].

Our work draws from parts of all the above mentioned systems to identify the presence of people in the scene. In our work, we integrate many different modules into a working system capable of tracking people and recognizing their gestures. We discuss these methods and the details of the various processing techniques that we employ in the next sections.

Section 2 of the paper discusses three different low-level tracking techniques and describes how to integrate them to obtain better tracking under various conditions. Section 4 gives an overview of the vision system on our mobile robot. Section 3 discusses how the results from Section 2 are used and applied in the context of gesture recognition. Future work and conclusions are discussed in Section 5.

# 2 Our Approach

We are interested in studying methods for robust tracking of people in a scene using a variety of techniques, each of which is good for certain situations. Towards this end, we first present specific techniques for low-level tracking of people. Then we describe how these techniques can be combined in tracking people in situations where the users are moving, the camera is moving and the scene is changing.

## 2.1 Tracking Techniques

**Color-based Tracking.** Color-based tracking using skin-color segmentation is a key step in many existing real-time people tracking systems. Our method for skin-color segmentation is based on the work of Yang and Waibel [22, 21] to extract an initial set of candidate faces. This segmentation technique relies on the fact that human skin colors cluster in a small region in a normalized color space, forming a well-defined skin-color distribution. Under certain lighting conditions, this skin-color distribution can be characterized by a multivariate normal Gaussian distribution. Using this distribution, we define a Gaussian color filter $c_i$ for any $i$-th pixel as:

$$c_i = \exp\left\{ (x_i - \hat{x})^T \, \Sigma^{-1} \, (x_i - \hat{x}) \right\}, \tag{1}$$

where $x_i$ is the color vector for the $i$-th image pixel, while $\hat{x}$ and $\Sigma$ are the mean color vector and covariance matrix of the skin color model. We have obtained sample images of human skin under various lighting conditions to define a skin color model and the related statistics. Using this color model we apply the above color filter to each pixel in the image and threshold the result. Median filtering is then performed on the resulting image to remove noise. This allows us to extract candidate faces and hands that are represented as segmented blobs in an image $I_c(x, y)$. An example of this image is shown in Figure 1(b).

**Motion Detection.** In addition to developing techniques for color-based tracking, we are also studying motion change detection in an image to determine the location of people in the scene. The motion detection process helps locate people in cases where color-tracking methods fail due to lack of acceptable lighting or when a user moves in front of objects that have similar colors as the skin (wooden furniture and doors cause many problems for color-based tracking, see Figure 3 (3)). Also, tracking people without static backgrounds or with a pan-tilt-zoom (PTZ) camera mounted on a mobile robot, (see Figure 6), requires us to develop methods that are robust to camera movements.

To supplement color-tracking with motion detection, we employ templates like motion change detection templates (MCT) and motion energy templates (MET). A very simple way of determining METs is *image differencing*. By subtracting pixel intensity levels from frames at time $t$ to $t + 1$ for an image sequence $I(x, y, t)$, we generate a binary image $B(x, y, t)$ where 1's indicate motion and 0's indicate no motion. If we know the time duration $T$ for any movement then we can calculate the MET $I_e(x, y, t)$ by:

$$I_e(t) = \bigcup_{i=0}^{T-1} B(t - i) \tag{2}$$

$T$ is an important parameter in this computation and in essence segments the sequence in time. We have arbitrarily chosen to compute $I_e$ for 25 frames. This value of $T$ works well for the method that we use to combine the information from different processes for tracking.

Using METs we can compute motion history templates (MHTs), $I_h(x, y, t)$. In MHTs pixel intensity accounts for the temporal history of motion at that point and is computed as follows ($x$, $y$, omitted for clarity):

$$I_h(t) = \begin{cases} T & B(t) = 1 \\ \max(0, I_h(t-1) - 1) & \text{otherwise} \end{cases} \tag{3}$$

The resulting template is a scalar valued image where the most recently moving pixels are brighter. Similar templates are used by Bobick and Davis [2] for recognition of actions in scenes. We also use our templates for recognition of gestures as discussed later. Examples of the motion energy and motion history images are shown in Figure 1(c) and (d).

**Stereo Estimation.** In addition to using color and motion segmentation to aid with tracking people in an environment, we are also using depth. Using depth information, we can estimate where the user is in the scene, which in turn, aids with tracking. Stereo information is really useful in scenarios where the camera is also moving in addition to the user, and there is variation in lighting. Depth information can also be used to foveate a PTZ camera closer to the user to acquire higher resolution imagery for further analysis.

Our first stereo estimation method relies on a commercially available stereo system from SRI [8]. This system correlates and matches on similar blocks to compute dense stereo images. We mounted this system on a PTZ camera. The PTZ camera is used to acquire regular color imagery and the stereo system provides depth. At present, the calibration between the stereo view and the monocular view is done manually so as to align all the objects in the scene correctly. We have also mounted this entire camera system on our robot as shown in Figure 6. The output from the stereo system is shown in Figure 1(e). This gray-scale depth image, $I_d$, can be used for tracking. However, this depth image is quite noisy and therefore is not sufficient for robust tracking of a person, a limitation addressed in the next section.

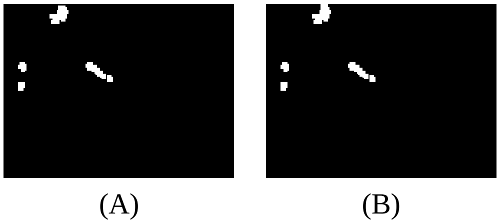
## 2.2 Merging information from different processes: Conditional Dilation.

In complex and dynamic environments we need to combine different processes to yield appropriate tracking of the users in the scene. We have discussed processes of color tracking, motion measurement, and depth estimation to yield color segmentation $I_c$, image motion $I_e$, and depth image $I_d$. The next step is to merge the different processes. A simple and obvious operation is the logical AND operator at each location in the image. The AND operation, however, will not represent moving objects very well. The reason for this is that $I_c$ is the upper bound on the region in which the moving skin-color objects can be contained. On the other hand, processes yielding $I_e$ and $I_d$ represent the lower bound on possible regions where the moving object can move to. The AND operator yields the minimum area of the two processes and therefore when the motion is small, the resulting image will not carry any information about the moving objects.

As opposed to just combining the information from the two processes, we segment moving skin-color objects $I_c$ by using information from either $I_d$ or $I_e$. The motion and depth processes serve as constraint regions on the segmentation of the color objects as the move. To achieve this type of segmentation, we use *conditional dilation*. Using this concept we define that *if* the moving skin-color objects can not occupy certain parts of an image due to color and motion constraints, *then* dilations of such objects must not intrude into the violated area. This is mathematically expressed as follows:

Let $\acute{I} = I_e \cap I_c$ be the image constructed by intersecting motion energy and skin color matching images, and let $S_e$ be the structuring element to be used in the dilation (We use a 3x3, origin centered full kernel as the $S_e$),

$$\acute{I}_{new} = (\acute{I} \oplus S_e) \cap I_c. \tag{4}$$



(A) (B)

Figure 2: *Using conditional dilation to combine two measurements. The resulting images show location of the head and the two hands. (A) Color and motion energy combined. (B) Color and stereo depth map combined. These are very similar in this instance. This similarity vanishes when the camera and subject both move as shown by later examples.*

Here $\acute{I}_{new}$ is the new segmented image of $\acute{I}$ and it can only be the same or better than $\acute{I}$ since it is bounded by $I_c$. $\acute{I}$ is computed iteratively to obtain a better regions of moving color objects. This merging process ideally continues until $\acute{I}_{new} = \acute{I}$ to get the maximum area of such objects. Performing these merging steps can be computationally expensive, but in our experiments, we have found that 4 to 5 iterations are sufficient and yield good results. The number of these iterations increases if we decrease the number of frames that we compute the motion energy over (Equation 2).

In the above formulation, we use $I_e$ to define the lower bound. $I_d$ can also be used as a lower bound in the cases where the camera motion is large in the same manner as $I_e$ is used in the following formulation.

In case of a moving camera the segmented regions of motions in $I_e$ are large and the resulting regions of movement are larger or equivalent to $I_c$. When a PTZ camera is tracking a person and when the person is moving slowly, the camera moves only when the person is moving and stops when and if it relocates the human again. This results in a lot of motion energy than in the case when the camera is still. Under these conditions, the conditional dilation of $I_c$ and $I_e$ still works, especially if the tracking of the PTZ reaquires the person before $T$ frames.

However, combination of $I_c$ and $I_e$ fails when the camera is moving fast and the PTZ camera fails to keep the person in the center. In this case, the depth image $I_d$ from the stereo system is combined with the skin color image $I_c$ using conditional dilation. The stereo depth image alone is not sufficient because it is noisy and blobs close to each other tend to be merged into one *i.e.,* the hands and the body of the person appear as one blob. Figure 2 shows the result of conditional dilation of the above two types for the images in Figure 1. Figure 3 shows results of conditional dilation on the color and motion processes (row 1), how this method works even when there is skin color in the background (row 2), how it fails when there is camera motion (row 3) and how a stereo depth image can be used by dilating with the color process to achieve reliable tracking for images from row 3 (row 4).

## 3 Gesture Recognition

Once we have found the position of the person in the scene we can try to recognize some of his gestures or activities (see section 4 for details on how the initialization of the vision system is done). We were especially interested in applying this for a human-robot interaction where the human will perform a gesture and the robot will react accordingly. We are using two different methods for gesture recognition.

In our first method, for each gesture that we want the robot to recognize we create a library of motion energy

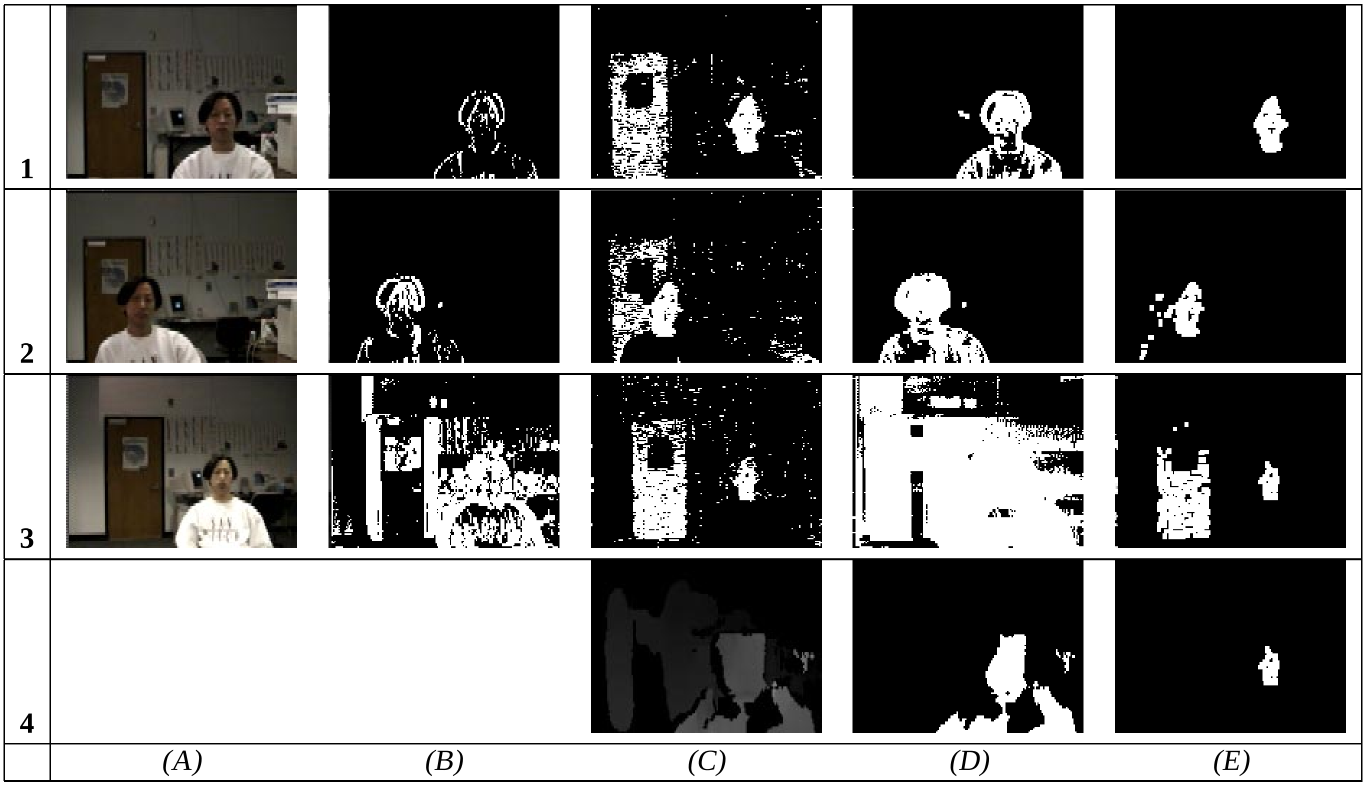


Figure 3: *Tracking a face using color, motion, and stereo processes. Row 1 shows a simple case of combining color and motion processes, Row 2 shows a case when there is a skin color in the background (a door), Row 3 shows an instance where conditional dilation of color and motion fails as the camera moves, and Row 4 shows that conditional dilation of color and depth processes succeeds in this instance. (A) Shows the original image, (B) shows motion difference between 2 frames, (C) shows the output of the color process, (D) shows the binary motion image, note the image in row 3. (E) shows the segmented/tracked region. In row 4, (C) is a grey-level depth image and (D) is the thresholded depth image.*

profiles [2] and train an artificial neural network (NN) to distinguish between these profiles.[1] The NN is trained off-line, however, the recognition is performed in real time (see Figure 4 for 3 of the 6 gestures that were trained for, each gesture had 15 instances).

In order for this technique to work, the motion energy profiles of the different gestures should be distinct enough. Also the directions of the gestures can not be obtained using motion energy profiles alone. We have tried using motion history profiles instead, but this requires a lot more training data to accommodate for the variations in the speed with which each gesture is performed.

For our second method, we use the centroids of the three blobs (head, left, and right hand) to form six dimensional vectors for each frame for the duration of the gesture. This was done 15 times for each gesture (10 used for training and 5 for recognition) and a hidden Markov model (HMM) [13, 16] based recognizer was trained on the ten instances. Recognition was achieved by matching to a corresponding HMM that produced the

[1] The images were reduced in size before the NN was trained to reduce complexity.

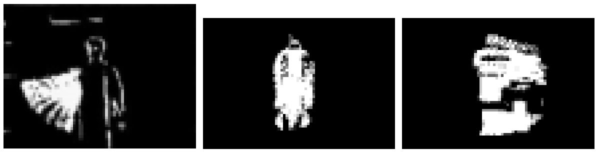


Figure 4: *3 examples of motion energy profiles used for training NN used for gesture recognition.*

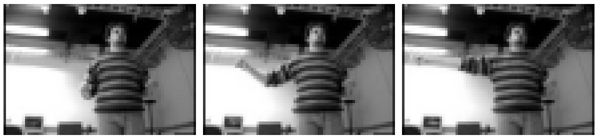


Figure 5: *3 frames captured during the beginning, middle, and end of a pointing gesture used for training an HMM recognizer.*

best score.

We trained six HMM's and out of the 30 tests sequences (6 gestures x 5 instances), 29 were recognized correctly and one was not recognized. For the one that was not recognized, the probability of each HMM generating it was below the threshold and was ignored. The HMM testing and training were done offline using Matlab code. We are currently integrating the HMM recognizer with the rest of the system to make it recognize gestures in real time.

## 4 Vision System on a Mobile Robot

Tracking people from a mobile robot is a much more challenging problem as the scene is much more dynamic because of both motion of the camera and that of the user. We have used the methods described so far in this paper on our robot shown in Figure 6. The images in Figure 1 are captured from the vision system that is being used with the robot.

The vision system initializes by segmenting out skin-colored regions. By doing simple constraints, the three blobs are labeled as hands and a face. When the user moves, the conditional dilation of color and motion processes takes over and allows for accurate tracking, even in front of objects that may have similar color features as human skin. When the robot starts moving, the vision system can still segment out moving skin color objects in the case of simple background. As discussed earlier, the stereo system will aid in tracking of large motions. We demonstrated an earlier version of this system at the AAAI-98 Robot Exhibition.

## 5 Conclusions and Future Work

In this paper we present our work on perception of people by a mobile robot. This is a significant step towards building a personal robotic agent; an agent that is aware of the users, and interacts with the users by interpreting gestures and speech. Towards this goal, we present techniques for reliable tracking of people in complex, dynamic and changing environments. We have show that standard techniques for skin-color tracking, motion detection and depth estimation are good for tracking people under specific conditions, however, are limited in dynamic and changing environments. We present a method based on conditional dilation to allow for merging

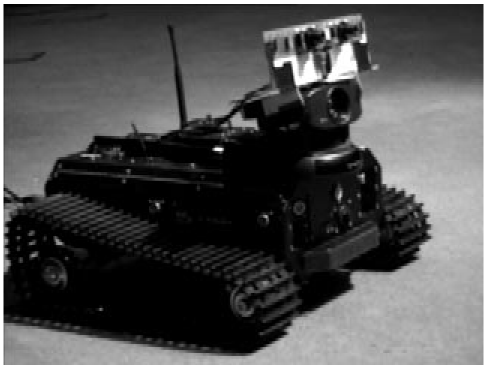


Figure 6: *The robot mounted with a PTZ camera and the SRI stereo system.*

of color, motion and depth measurements. We show the importance of this type of multiple process merging for tracking people.

We show the validity of these methods by applying them to a robotic vision system. This system can track a person under varying lighting conditions, with large camera and background movements and recognize static gestures. These are important steps in our effort to make a robot that can be aware of the user and recognize the user's gestures and activities.

For the future, we are interested in using other measurement processes and attempt to recognize gestures from the robotic "eyes" as the robot and the user continuously move.

# References

[1] A. Azarbayejani and A. Pentland. Real-time self calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proceedings of the International Conference on Pattern Recognition 1996*, Vienna, Austria, 1996.

[2] A. F. Bobick and J. W. Davis. An apearance-based representation of action. In *Proceedings of International Conference on Pattern Recognition 1996*, August 1996.

[3] J. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Press, 1997.

[4] T. Darrell, I. Essa, and A. Pentland. Attention-driven expression and gesture analysis in an interactive environment. In *International Workshop on Automatic Face and Gesture Recogntion*, pages 135–140, Zurich, Switzerland, 1995. Editor, M. Bichsel.

[5] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Multi-modal person detection and identification for interactive systems. In *Proceeding of Computer Vision and Pattern Recognition Conference*. IEEE Computer Society Press, 1998.

[6] I. Haritaoglu, D. Harwood, and L. Davis. W4s: A real time system for detecting and tracking people in 2.5 d. In *Eurepean Conference on Computer Vision*, 1998.

[7] T.S. Jebara and A.P. Pentland. Parametrized structure from motion for 3d adaptive feedback tracking of faces. In *Proceedings of Computer Vision and Pattern Recognition Conference*, pages 144–150, 1997.

[8] K. Konolige. Small vision systems: Hardware and implementation. In *Proceedings of Eight International Symposium on Robotics*, Hayama, Japan, November 1997.

[9] D. Kortenkamp, E. Huber, and R. P. Bonasso. Recognizing and interpreting gestures on a mobile robot. In *Proceedings of AAAI-96*, pages 915–921, 1996.

[10] Pattie Maes. ALIVE: an artificial life interactive video environment. In *ACM SIGGRAPH Visual Proceedings*, page 189, MIT Media Laboratory, 1993.

[11] D. Murray and A. Basu. Motion tracking with and active camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 16(5):449–459, 1994.

[12] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition Conference*, pages 84–91. IEEE Computer Society, 1994.

[13] L Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[14] J.M. Rehg, M. Loughlin, and K. Waters. Vision for a smart kiosk. In *Proceedings of Computer Vision and Pattern Recognition Conference*, pages 690–696, 1997.

[15] H.A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *Proceedings of Computer Vision and Pattern Recognition*, 1998.

[16] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *Proceedings of International Workshop on Automatic Face and Gesture Recogntion*, Zurich, Switzerland, 1995.

[17] S. Stillman, R. Tanawongsuwan, and I. Essa. A system for tracking and recognizing multiple people with multiple cameras. In *Proceedings of Audio and Vision-based Person Authentication 1998 (To Appear)*, Washington, DC, USA, March 1999.

[18] K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report 1521, MIT AI Laboratory, 1995. http://www.ai.mit.edu/.

[19] S. Waldherr, S. Thrun, R. Romero, and D. Margaritis. Template-based recognition of pose and motion gestures on a mobile robot. In *Proceedings of Annual AAAI Conference 1998*, 1998.

[20] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[21] J. Yang, W. Lu, and A. Waibel. A real time face tracker. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, volume 2, pages 687–694, 1998.

[22] J. Yang and A. Waibel. Skin-color modeling and adaptation. In *Proceedings of IEEE Workshop on Applications of Computer Vision*, pages 142–147. IEEE Computer Society Press, 1996.