Multiple-model Based Human Tracking

Ying REN, Chin-Seng CHUA, Yeong-Khing HO
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798
P144775376@ntu.edu.sg

Abstract

This paper describes a system for real-time multiple human tracking in an unconstrained environment using a single color camera. The use of multiple cues (motion, figure and color) is able to constrain, assist and complement each other in a complicated tracking process and environment. The background is retrieved and adapted dynamically using a pixel-wise statistical model while the moving regions are detected, without obtaining the background in advance. Figure and color models are integrated with the motion information to extract and track the geometric and visual features of the human body across a sequence of images. Figure model is able to discriminate the human body from other moving objects and define a stable region as the geometric feature. Color model of the human's clothes is established automatically and used as one of the cues in the tracking process with multiple persons. The results based on the two models are combined to make the decision in the object correspondence procedure. Experimental results have shown the effectiveness and efficacy of the system in a real-time environment.

Keywords: *Human tracking, model-based, background retrieval, motion segmentation.*

1. Introduction

Robust human tracking is a prerequisite for automatic security surveillance system, face recognition, motion analysis, advanced user interface, model-based coding and virtual reality. There are a considerable amount of research efforts [11, 1, 4, 2, 5, 9] about human detection and tracking in the past few years. "Pfinder" [11] detects and tracks a single human body in a static background using a set of 2D models, which are described by the spatial (x, y) and color (Y, U, V) Gaussian distributions over the pixels they consist of. These models typically correspond to the person's hands, head, feet, shirt, and pants. The background is described by the Gaussian distribution pixel by pixel in terms of color values, and it is trained in advance without any foreground occlusion. [1, 4] are multiple human tracking systems that focus on the movement of the whole human body using a monochromatic camera. Figure model of the human body is used to recognize and track parts of body. On the other hand, [5, 9] proposed a color-based human tracking systems. Color does not change significantly with camera movement, and it can provide some robustness with respect to occlusion. However this method requires appropriate color initialization, where the object's color has to be defined manually. In addition, the constraint of distinguishing color between the objects and the background is required. Each of the above single-cue(motion, figure or color) based human tracking system tends to work well over a limited range of conditions, but often fail when exposed to a "real world" environment.

In this paper, we describe a multiple human tracking system with a static color camera. This system integrates multiple cues of motion, figure and color to handle the occlusion, abrupt change of the direction of movement and other complicated issues in the tracking process. The background is retrieved and adapted dynamically using a pixel-wise statistical model while the moving regions are detected frame by frame and adapted with the changing environment. A simple figure model is adopted to discriminate the human being from other moving objects and describe the moving region that corresponds to the human body. Skin color model is used to detect a person and automatically locate a sample region that can be used for the training of the clothes color model. The clothes color model is established and trained on-line when the person appears and used as one of the cues in the tracking process. The results of color model and figure model are integrated to establish object correspondences across the sequence of continuous images, especially in complicated situations when the singular use of any of the cues fails. Our work is similar to that work of Darrell *et al.* [2], which uses stereo, skin color and facial pattern to detect and track a head model. It needs special hardware for the range data and it can only track the head with frontal gesture.

The remainder of this paper is arranged as follows. Section 2 describes the technique for moving region detection. Section 3 introduces the figure model and geometric feature extraction. Section 4 reviews the color model, color model initialization, prediction and adaptation. Section 5 describes a tracking system based on motion, figure and color cues. Finally, conclusions are drawn in section 6.

2. Moving Region Detection

The initial stage of tracking is the detection of the moving objects. Generally, there are three motion-based approaches for moving object detection in a still background, namely, temporal difference [1], optical flow [8] and background subtraction [4]. The background subtraction approach provides the best results when compared with the other two methods, in terms of the computational requirements and the information they can provide. However there may be situations where no clean background is available in advance. For a continuous tracking system, the background is not always the same. It changes with illumination and scene content. An assumed invariable background with simple threshold for the foreground detection is thus not suitable.

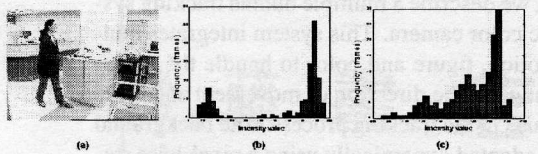

Figure 1. (a) The No.75 frame of a image sequence with image resolution of 320×240, which depicts a person roaming in the laboratory. (b) the distribution of intensity values for pixel (80, 80) over 200 frames, where the background is covered repeatedly by clothes of the person. (c) the distribution of intensity values for pixel (90, 190) over 200 frames, where the background is covered repeatedly by the shoes, pants and shadow.

Here we do not assume that the background is obtained *a priori*. Each pixel is modeled as an independent statistical process [3], a mixture of Gaussians. Each Gaussian corresponds to the distribution of background, shadow, and different moving objects covering this pixel over time. Referring to Figure 1, (a) shows one frame of a sequential of 200 images, with resolution of 320×240, which depict a person roaming in the laboratory. (b) shows the distribution of intensity values for pixel (80, 80) over 200 frames, where the background is covered repeatedly by the clothes of that person. The distribution with a higher frequency on the right most is of the background, the left most one comes from the images of the person's clothes. (c) shows the distribution of intensity values for pixel (90, 190) over 200 frames, where the background is covered repeatedly by the shoes, pants and shadow. The distribution with a higher frequency on the right most is of the background, and the other distribution corresponds to the person's shoes, pants and shadow moving past the pixel along the image sequence. The distributions are different from pixel to pixel and can be fitted by multiple Gaussians which compose the Gaussian mixture model of each pixel. Each pixel is represented by a color vector[1], $\mathbf{V} = [H, S, I]^T$. Assuming there are c Gaussian distributions for a pixel (x, y), the model is parameterized by the parameter set $\Theta_i = \{\omega_i, \mu_i, \Sigma_i : i \in (1, \ldots, c)\}$, and the Gaussian mixture is

$$p(\mathbf{V}) = \sum_{i=1}^{c} \omega_i N(\mathbf{V}; \mu_i, \Sigma_i) \tag{1}$$

and

$$\sum_{i=1}^{c} \omega_i = 1 \tag{2}$$

where c is the number of the Gaussian distributions, ω_i, μ_i, Σ_i are the weight, mean vector and covariance matrix of the i^{th} Gaussian in the mixture model respectively. The probability density function of the i^{th} Gaussian $N(\mathbf{V}; \mu_i, \Sigma_i)$ is described as

$$N(\mathbf{V}; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(V-\mu_i)^T \Sigma_i^{-1} (V-\mu_i)} \tag{3}$$

where n is the dimension of the vector $\mathbf{V}$.

The parameters Θ_i of the Gaussian mixture model together with the number c of the Gaussian distributions can be obtained on-line using the first few frames. A divisive vector quantization [10] is chosen to approximately estimate the parameters of the model for each pixel instead of EM algorithm [3] in consideration of computational requirement. The input data $\mathbf{V}(t)$ is compared with each Θ_i, the best-matching Gaussian with parameter Θ_w is adapted using a constant learning rate ε[12]. and the parameters $\{\Theta_i, i \neq w\}$ of other Gaussians are left unchanged except the weights. Each weight of the Gaussian should be adapted to meet the demand of Equation (2).

$$\omega_i(t) = (1 - \varepsilon)\omega_i(t-1) + \varepsilon \cdot M_i(t) \tag{4}$$

$$\mu_i(t) = (1 - \varepsilon)\mu_i(t-1) + \varepsilon \mathbf{V}(t) \tag{5}$$

[1] In consideration of the changing illumination and shadow, we choose the HSI color space and use the component hue and saturation of the HSI in the procedure of moving region segmentation.

$$\Sigma_i(t) = (1-\varepsilon)\Sigma_i(t-1) + \varepsilon(\mathbf{V}(t) - \mu_i(t))^T(\mathbf{V}(t) - \mu_i(t)) \quad (6)$$

where, $M_i(t)$ is a binary value describing the winning Gaussian at time t. For the winning Gaussian, $M_i(t) = 1$, otherwise, $M_i(t) = 0$. ε is the learning rate and $0 < \varepsilon < 1$.

If there is no Gaussian that can best-match the input data or the variance of the input data is greater than a threshold, a new Gaussian is added with higher variance, and the mean equals the vector value of that input data. The Gaussian with the weight lower than a threshold in some duration will be discarded. The weights of all Gaussians are changed accordingly.

According to Friedman and Russell [3], the variance of the background should be smaller than others, and the weight should be stable in the sequence. Here we select the distribution with the highest $\omega_i/|\Sigma_i|$ as the background. The pixel value that does not belong to the background distribution will be regarded as the foreground.

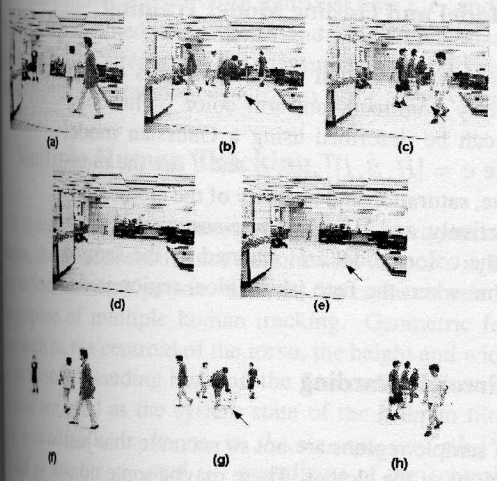

Figure 2. Background retrieval, adaptation and moving region segmentation.

Figure 2 shows the experimental results of background retrieval, adaptation and moving region segmentation. (a)-(c) are three images of a sequential of 600 images. While several persons are roaming in the laboratory, a box (pointed by an arrow) is brought into the scene and becomes part of the background. As shown in (d)-no box, and (e)-presence of box, the background is retrieved and adapted using the tracking procedure. At the same time, foreground regions are detected. (f)-(h) show the corresponding segmented foreground regions(compare with (a)-(c)).

3 Figure Model and Geometric Feature Extraction

The result of the foreground detection is a binary image. In practice, this binary segmentation is not perfect. It may be attributed to several reasons: Firstly, if the colors of the foreground and the background are similar, the foreground will be regarded as a part of the background and cannot be detected. Secondly, there will be some corruption due to sensor noise. Morphological operations such as erosion and dilation with structure elements are applied to remove small regions in the background and fill in small holes in the foreground region. Connected component labeling algorithm [6] is used to group pixels according to their spatial relationship and label the different regions.

The human body is an articulated object. Due to the periodic motion of the upper and lower limbs, the whole region that represents the human body will not be stable in continuous images. We regard the torso region as more stable than the other parts of the human body. The centroid and bounding box of the torso are extracted as geometric features for the purpose of tracking. Figure 3(a) describes the simple figure model of the human body. As a coarse discrimination criterion, the ratio of the height to width of the whole body should be greater than 2. The centroid and the bounding box are represented in terms of the first and second order moments [7]. However, the bounding box of the upper body should be further refined by analyzing the statistics of the upper region's vertical projection histogram due to the swing of upper limbs, as illustrated in Figure 3(b). The top row of Figure 4 shows the the bounding boxes of the human torso without refinement. The width of the bounding box and the centroid of the torso are affected by the upper swing limbs. The bottom row of Figure 4 shows the bounding box after refinement.

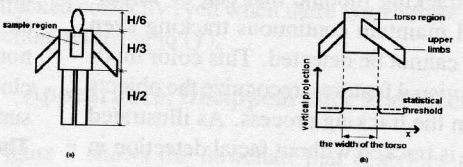

Figure 3. The frontal figure model of human body and torso region refinement.

4 Color Model and Visual Feature Extraction

Color offers many advantages over geometric information for those problems such as robustness under partial oc-

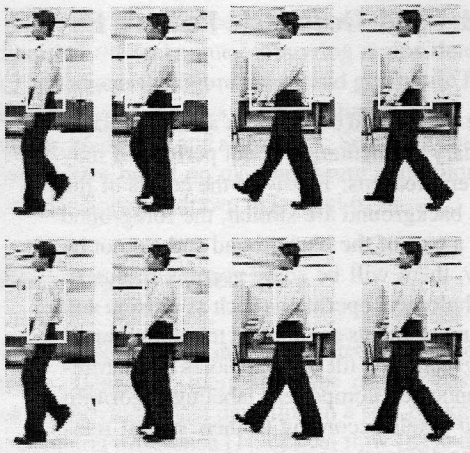

Figure 4. Top row: the bounding boxes of the torso without refinement. Bottom row: the bounding boxes after refinement.

clusion, rotation, scale and resolution changes. In addition, color based tracking can achieve fast computational speed. So far, the color based human tracking system can be categorized into two classes. In the first class, skin color model is used to detect the facial region frame by frame and localize the human head for the purpose of tracking [2]. When face cannot be found even when the person exists (for example, the person turns around and the facial color, pattern cannot be detected), these systems will fail. In the second class, the color of the person's clothes is used as the tracking cue [5, 9]. These systems usually require appropriate color initialization where the object's color has to be defined manually, and the constraint of distinguishing color between the objects and the background is required. Here we propose a color based human tracking module that can be initialized automatically and maintain continuous tracking even when the facial region cannot be detected. This color model is used to extract the visual features, recognize the objects and keep on tracking in the tracking process. As illustrated in Figure 5, the person is tracked without facial detection in every frame. When self-occlusion happens(Figure 5(b, c)), it can keep on tracking.

Comparing with the skin color, the clothes color of the person is more distinguishing for the purpose of tracking. By the motion constraint, confusion between the object and background is reduced. When the person first enters the scene, the human face is detected using the skin color model and the facial region is constrained within the upper foreground region. (This is a valid assumption for most surveillance systems, where the frontal view is usually available when a person enters the scene.) The regions with skin color in the background and the regions of the hand will not be taken into consideration. Sample region of the clothes is located by using the structure of the human body, as defined in Figure 3(a). The sample regions have the same size with the face regions and located just below the face. The data in the sample regions are used to train the clothes color model. Once trained, the face detection is not required in subsequent frames and the clothes color model is used as visual feature to recognize and locate the tracking object.

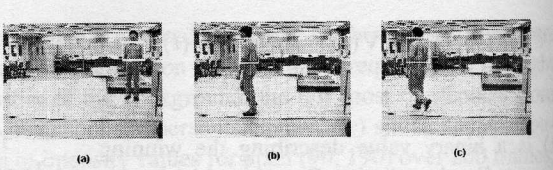

Figure 5. The color model based human tracking with the motion constraint.

4.1 Color Model and On-line Model Training

Here we have an assumption that the person's clothes have a sufficiently dominant, uniform color. The color of the clothes can be described using a Gaussian model $N(\mathbf{c}, \Sigma)$, where $\mathbf{c} = [\bar{H}, \bar{S}, \bar{I}]^T$, $\bar{H}$, $\bar{S}$ and $\bar{I}$ are the mean values of the hue, saturation and intensity of the clothes color model respectively, and Σ is the covariance matrix. The parameters of the color model are obtained by the sample data of the frame when the face (skin color) region is detected.

4.2 Large Error Discarding

The defined sample regions are not so accurate that just fall into the regions of the clothes. There maybe some non-clothes regions included in the sample regions and these non-clothes regions are small enough compared with the clothes regions. For computational consideration, we assume that the hue, saturation and intensity are independent. The thresholds,$\bar{H} \pm 3\sigma_H$,$\bar{S} \pm 3\sigma_S$ and $\bar{I} \pm 3\sigma_I$, are used to discard the sample data with large errors. And then the model parameters are recomputed.

4.3 Color Model Prediction and Adaptation

Hue and saturation are less influenced by illumination changing. So the skin color model is establish in HS color space. In the clothes color model, intensity is added as one of the parameters for discriminating the color in some special situations. For example, white and black have the same hue and saturation but definite opposite intensity. Due to the

Illumination change and the mutual reflection of the objects, the parameters of the clothes color model will be changed continuously in the image sequence, especially when the objects are moving closed to the camera. Prediction and adaptation of the model's parameters are required in the color based tracking process. The approximate parameters' prediction and adaptation are as follows,

Prediction:

$$\hat{\mathbf{c}}_{t|t-1} = \bar{\mathbf{c}}_{t-1} + (\bar{\mathbf{c}}_{t-1} - \bar{\mathbf{c}}_{t-2}) \quad (7)$$

$$\hat{\Sigma}_{t|t-1} = \bar{\Sigma}_{t-1} + (\bar{\Sigma}_{t-1} - \bar{\Sigma}_{t-2}) \quad (8)$$

Adaptation:

$$\bar{\mathbf{c}}_t = (1-\alpha)\hat{\mathbf{c}}_{t|t-1} + \alpha \mathbf{c}_t \quad (9)$$

$$\bar{\Sigma}_t = (1-\alpha)\hat{\Sigma}_{t|t-1} + \alpha \Sigma_t \quad (10)$$

where, $\hat{\mathbf{c}}_{t|t-1}$ and $\hat{\Sigma}_{t|t-1}$ represent the estimated mean vector and estimated covariance matrix at time t, $\bar{\mathbf{c}}_t$ and $\bar{\Sigma}_t$ represent the adapted mean vector and adapted covariance matrix, $\mathbf{c}_t$ and Σ_t represent the measurement mean vector and measurement covariance matrix, α is the adaptive rate.

5 Multiple Human Tracking

In order to track the human body across a sequence of images, the system should not only be able to locate it, but also to find the same person. Kalman filters are chosen for the purpose of multiple human tracking. Geometric features such as the centroid of the torso, the height and width of the torso's bounding box and the first order derivatives of them are used as the system state of the Kalman filter. It is defined as $\mathbf{s}_t = [x_t, y_t, w_t, h_t, vx_t, vy_t, vw_t, vh_t]^T$, where x_t and y_t are the $x-$ and $y-$ coordinate of the centroid; w_t and h_t are the width and height of the torso's bounding box; vx_t, vy_t, vw_t, vh_t are their first order derivatives respectively.

5.1 Object Correspondence

In the object correspondence stage, The results of the figure model and color model are integrated to decide the location and correspondence of objects between the two successive images.

Mahalanobis distance D_t is used to determine the correspondence of geometric feature being tracked. The measurement $\mathbf{z}_t$ of the object in the current frame is Gaussian distributed about its predicted value $\hat{\mathbf{s}}_{t|t-1}$,

$$D_t = [\mathbf{z}_t - \hat{\mathbf{s}}_{t|t-1}]^T \Lambda_t^{-1} [\mathbf{z}_t - \hat{\mathbf{s}}_{t|t-1}] \quad (11)$$

where, Λ_t represents the covariance of the error $[\mathbf{z}_t - \hat{\mathbf{s}}_{t|t-1}]$. The locus of points of the given Mahalanobis distance is a hyper-ellipsoid. The center of this hyper-ellipsoid is the prediction $\hat{\mathbf{s}}_{t|t-1}$. If the measurement of current frame falls into the hyper-ellipsoid, it is regarded as one of the candidates of the object being tracked.

The centroid $\mathbf{C}_C$ of the recognized object based on the color model should have the same coordinates with the centroid $\mathbf{C}_G$ of the torso region based on the figure model. If the inequalities

$$\mathbf{C}_C \leq \mathbf{C}_G \pm \delta \quad (12)$$

$$D_t \leq \gamma \quad (13)$$

are satisfied, the correspondence between the region in the current frame and that being tracked is uniquely decided. δ is the accepted error between $\mathbf{C}_C$ and $\mathbf{C}_G$. γ can be obtained from χ^2 distribution table.

Single model based tracking is not robust in complicated environment. Tracking based only on the color model will fail when the moving objects are of the same color. Tracking based only on figure model will have problems as well. For example, when a person suddenly changes direction, the motion estimation will provide a wrong prediction of the motion. Using the geometric features based on figure model to decide the object correspondence will result in errors. In this case, the equation (12) is not satisfied. The coordinate of the color model based centroid is more reliable, so that the object is selected as the correspondent object of the one being tracked. Referring to Figure 6, the top row shows the tracking results based on the figure model. When persons change their motion direction suddenly, the person with the black bounding box is lost by the tracker, and the person with the white bounding box is tracked incorrectly, as shown in (c). The bottom row shows the tracking results based on the color model with the motion constraint. Persons with black and white bounding box are tracked correctly even when there is an abrupt change in direction, referring to (f).

5.2 Appearance, Disappearance and Occlusion

Figure 7 shows the flowchart that describes the tracking strategy including the situation of appearance and disappearance. Occlusion can be considered as the special situation of the appearance and disappearance, the object disappears for a while and then reappears. This object will not be discarded at once. It will be kept in the tracking list for m frames. After it reappears, the object is kept on tracking.

Figure 8 shows the multiple human tracking based on motion, figure and color models. Colors of the bounding box show the tracking results. On a Pentium III-450, with an image resolution of 320 × 240, the speed of the integrated system is about 5 frames/second.

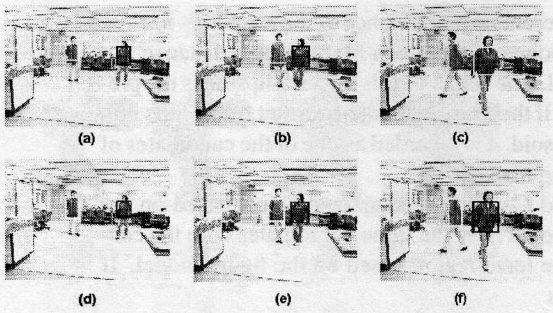

Figure 6. Multiple-model based human tracking in the condition of changing direction suddenly.

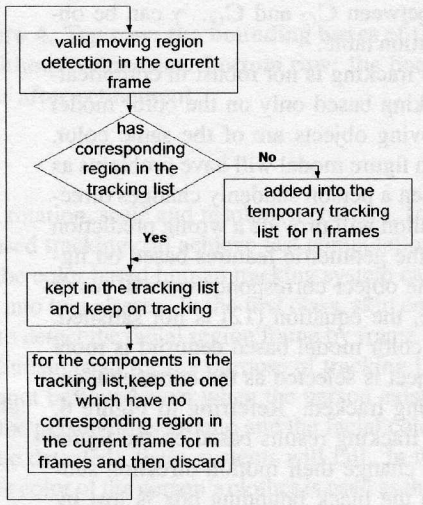

Figure 7. Flowchart of the tracking strategy

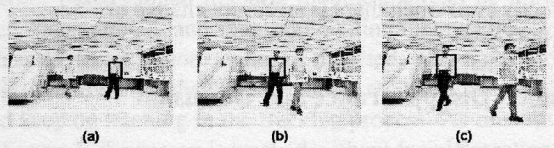

Figure 8. The multiple human tracking based on motion and model.

6 Conclusions

This paper presents a real-time human tracking system based on multiple-model. Figure model and color model are applied to extract the geometric and visual features of the human body on the basis of background retrieval, adaptation and foreground region segmentation. Kalman filter is utilized for the tracking of multiple human in the sequence of images. Color model of the clothes is established automatically and used as one of the cues in the tracking process. The results based on the two models are integrated to make the decision in the object correspondence procedure. Experimental results show the robustness of the system in complicated situations and its real-time performance.

References

[1] Q. Cai and J. K. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. *Proceedings of the 6th International Conference on Computer Vision*, pages 356–362, 1998.

[2] T. Darell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 601–608, 1998.

[3] N. Friedman and S. Russell. Image segmentation in video sequences: a probabilistic approach. *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 1–3, 1997.

[4] I. Haritaoglu, L. S. Davis, and D. Harwood. W4: who? when? where? what? a real time system for detecting and tracking people. *Proceedings of the 3th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 222–227, 1998.

[5] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 257–260, 1997.

[6] B. K. P. Horn. Robot vision. *The MIT Press*, 1986.

[7] A. K. Jain. Fundamentals of digital image processing. *Prentice Hall*, 1989.

[8] M.Etoh and Y. Shirai. Segmentation and 2d motion estimation by region fragments. *Proceedings of International Conference on Computer Vision*, pages 192–199, 1993.

[9] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using color. *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 228–233, 1998.

[10] C. W. Therrien. Decision estimation and classification. *John wiley and sons*, 1989.

[11] C. R. 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.

[12] P. Young. Recursive estimation and time-series analysis. *Springer-Verlag*, 1984.