JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 26, 2267-2281 (2010)

Particle Filter-based Multi-part Human Tracking with Failure Adjustment in Video Sequences

SAN-LUNG ZHAO AND HSI-JIAN LEE*
Department of Computer Science
National Chiao Tung University
Hsinchu, 300 Taiwan
E-mail: slzhao@csie.nctu.edu.tw
**Department of Medical Informatics*
Tzu Chi University
Hualien, 970 Taiwan
E-mail: hjlee@mail.tcu.edu.tw

The study presents a human tracking system in video sequences. To track the target, we first detect humans in a video according to a Gaussian background model. We then track the humans by using color histogram as the features and using particle filters as the tracking kernel. Since a human is not a rigid object, his appearance may be greatly affected by his motion. In our applications, human bodies imaged are generally large. We decompose each human body into three parts: head, torso, and hip-leg, represent them by three shrunk rectangles, and track them by particle filters. In this way we can reduce possible tracking failures by checking the interrelationship among these three parts. We use support vector machines (SVM) to detect tracking failures and abnormal body parts since the abnormal situations are very diversified and cannot be easily encoded in rules. If a single part is abnormal, its position can be adjusted from the other parts and tracked using the system dynamic model. If two or three parts are abnormal, we re-initialize the tracking process of the three parts around their predicted positions. By testing on 22 video clips from six scenes, the experimental results showed that our three-part tracking system with failure detection and correction can track correctly about 95% persons until the 105th frame. With respect to the body parts, our system has about 95%, 83%, and 91% tracking rates for head, torso, and hip-leg respectively until the 105th frame. The tracking rate of a human increase 20% comparing with that of the whole-body tracker. These rates show the effectiveness of the proposed system.

Keywords: human tracking, particle filter, support vector machine, tracking failure adjustment, multi-part tracking

1. INTRODUCTION

Human tracking is a fundamental and important step for many visual surveillance applications, such as security guard, patient care, and human-computer interaction. To track an object in a sequence of frames, we can model appearances of the object and then use the model to predict its position in the sequence. However, in a complex environment, detecting a target object using the appearance model in video sequences is not easy since the appearances of the object are variable due to occlusion, illumination variations, or orientation changes. In general, the movements of an object in consecutive frames are assumed smooth. Therefore, if we can locate the target object in several frames, the appearance model and movement model of the target object obtained from these frames can

Received August 20, 2008; revised November 11, 2008; accepted January 22, 2009.
Communicated by Tong-Yee Lee.

be used to track the object in the following frames.

In this study, we aim to create the trajectory of a human and predict his positions for safeguarding, that is, to detect an intruder approaching a building, an entry, wall or a designated place. Since a human is not a rigid object, his appearance might be greatly affected by his motion. We decomposed the human body into three parts: head, torso, and hip-leg, since the three parts usually have different appearances and can be distinguished. The images show that the colors of the head part contain mostly skin colors and hair colors, which are usually different from the colors of the other two parts. The colors of the torso and hip-leg parts consist mainly of those of the clothes, which may be similar. To separate the three parts, we have to use other features such as height ratios.

With respect to the features used, we adopted color histograms proposed by Perez *et al.* [1] and Nummiaro *et al.* [2] to model the appearances of the three body parts. In the initialization phase, we adopted the background subtraction method according to a Gaussian background model to extract a human and then extract the histograms of the body parts from the human region. However, when modeling the color histogram in the whole color space, histogram matching is time-consuming due to the high dimensional features used. The method proposed by Nummiaro *et al.* [2] quantized the color histogram into an $8 \times 8 \times 8$ or $8 \times 8 \times 4$ three-dimensional one. The method proposed by Perez *et al.* [1] modeled colors in HSV color space by two histograms. The intensity channel was modeled as a histogram and the other two channels as another two-dimensional histogram. The histograms were quantized into several bins to improve the speed and reduce the effect of noise. However, in these models, two objects with very few dissimilarities were not easily distinguished. In our research, we equalized the color histograms to improve the ability of discriminating the objects with similar color distributions.

For failure detection and adjustment, we will use a support vector machine (SVM) [3, 4] to distinguish abnormally and normally tracked body parts. The position of an abnormal body part will be adjusted according to its relative positions with the other body parts. If a single part was abnormal, we adjusted its position and used the system dynamic model to track the abnormal one. If two or three parts were abnormal, we re-initialized the tracking process of the three parts around their predicted positions.

The remainder of this paper is organized as follows. Section 2 gives related work for human detection and object tracking. Section 3 describes the method of particle filtering for object tracking. Section 4 describes the module of three-part human tracking and consistency checking, which uses four SVMs to check whether there is a tracking failure and adjust the failure part according to the relative positions among the three body parts. Section 5 gives experimental results and their analysis. Finally, section 6 presents conclusions and suggestions for future work.

2. RELATED WORK

In the last few decades, tracking objects or humans in video sequences has received much attention. Much research about the topic has been proposed and been reviewed in several survey papers [5-7]. Moeslund *et al.* [5] divided a general human tracking algorithm into two main phases: *figure-ground segmentation* and *temporal correspondences*. The former finds the target human in an image, and the latter associates the detected humans in consecutive frames to create temporal trajectories. In the following, related work

about these two phases will first be addressed. The methods for segmenting human bodies and correcting tracking failures will then be described.

The methods of *figure-ground segmentation* can be classified into five categories according to the used features. These categories include *background subtraction* [8, 9], *motion-based segmentation* [10], *depth-based segmentation* [11], *appearance-based segmentation* [1-3, 12, 13], and *shape-based segmentation* [14]. Background subtraction and motion-based segmentation methods find the differences between images to extract the target. The depth-based segmentation approach uses the positions of the target in three-dimensional space or in the ground plane to segment the target. The appearance-based segmentation approach became popular recently, since the approach is usually simple and fast. The approaches of shape-based and appearance-based segmentation are similar except that the former does not use the color content inside the object. Since the appearances of a tracking target may change with time, several researchers proposed methods to model and update the appearance model of the target person dynamically in consecutive images [2, 13]. Other researchers used classifiers such as SVM [3] and Adaboost [12] to model the appearance of target objects.

In the tracking phase, *temporal correspondence* aims to predict and update the states of the target person from the measurement and predicted state. To combine the predictions and measurements, Kalman filtering is a well-known method and has already been applied in many studies [8, 15, 16]. The Kalman-filter-based approaches are commonly used for tracking a target whose system dynamic model can be represented as a linear function and the noise as a Gaussian. Recently, particle filters have been proposed to construct a robust tracking framework that are neither limited to linear dynamic model nor Gaussian distributed noise [2, 17]. The method represents the state of a target object by a set of samples (particles) with weights. The weight of a sample is calculated by the figure-ground segmentation and the samples are generated by the importance sampling method so that the samples can represent the probability distributions of the target object's appearances. We adopt the particle filter in our system, since they can be applied in an appearance-based tracking system very effectively.

A human is not a rigid object and his appearance changes irregularly. Segmentation of human body parts in an image has already been proposed in several papers [18, 19]. Ioffe and Forsyth [19] decomposed the human body into nine distinctive segments. The method finds a person by constructing assemblies of body segments. The segments were consistent with the constraints on the appearance of a person that result from kinematic properties. Recently, body-parts-based human tracking in consecutive images has been proposed [3, 12, 20, 21]. Parts of these studies focused on precise decomposition of body parts for motion type or pose analysis. However, in general environments, it is difficult to decompose precisely body parts due to self occlusions and complex background scenes. The studies in [3, 21] proposed a detection-based tracking model to solve the occlusion problem. When multiple humans appeared in a frame, the detection model could not differentiate the body parts of the different persons.

When a person is tracked in consecutive frames, the figure-ground segmentation may fail, since the person may be occluded or other objects may have similar appearances with the target person. The first problem can be classified into occluded by other persons and occluded by background objects. To cope with the problem of inter-person occlusion, several researchers proposed to detect occlusion events and then used the sys-

tem dynamics to estimate the position of the occluded person [22, 23]. When a body part is occluded, the position of the person can still be tracked based on the other parts. Mohan *et al.* [3] extracted a human body by detecting four parts: the head, legs, left arm, and right arm, by four distinct quadratic support vector machines. After geometric constraints among these parts are confirmed, another support vector machine is used to classify the combination of the four parts as either a human or a non-human. Wu and Nevatia [21] used four detectors to detect head-shoulder, torso, legs, and full-body. They used a strong classifier to classify the body parts in images. When we track multiple humans, the classifier cannot be used to distinguish different persons, and their trajectories will easily be confused if no other approaches are adopted. In our research, we use an adaptive appearance model to track the body parts, even when multiple persons are tracked.

Apart from the occlusion events, a tracker may lose the tracking target when other objects have similar appearances. In general, a robust appearance model can be used to reduce the tracking failures, or the system dynamic model of the target person can be used to predict his position. However, the robust appearance model may be too complex to maintain efficiently. We will use the system dynamic model of the target person to track him, when a tracking failure is detected.

3. PARTICLE FILTER FOR OBJECT TRACKING

A tracking algorithm is usually composed of two procedures: prediction and update. In the prediction procedure, the system dynamic model of the target is used to predict the current state of the target from previous states. In the update procedure, current observations are used to adjust the predicted state of the target.

In this research, we adopt the color-based particle filter proposed by Nummiar *et al.* [2] to track the targets. In a particle filter, a target object is tracked by a set of weighted sample states (particles). In the prediction procedure, the samples are propagated into the next step according to the system dynamic model. The update procedure can be divided into two steps: particle weighting and particle selection. In the first step, the weight of a sample is calculated according to the target model, which models the observations of a target object and can be used to calculate the probability of a sample belonging to the target. In the second step, the Monte-Carlo method is used to re-sample the particles.

The target state used in this research is described as a vector $S = [X, Y, W, H, \dot{X}, \dot{Y}]^T$ where (X, Y) represents the center of the rectangle, (W, H) the size of the rectangle, and $(\dot{X}, \dot{Y})$ the velocity of the center. The system dynamic model is defined as a uniform velocity motion model. The particle weighting is identical to that described in [2]. To improve the discriminability between different objects, we propose an equalized color histogram as the observation model. In a particle filter, a large number of particles may cause low speed, but a small number of particle may cause low accuracy. To efficiently utilize the particles, we dynamically adjust the number of particles according to the entropy of particle weights. In the following, we will explain the observation model and adjustment of the number of particles.

3.1 Equalized Observation Model

In our application, we aim to track a human in consecutive color images. The color

histogram model [2] is robust against partial occlusion, non-rigidity, and rotation. However, in our application, the region of a tracking target may be small. To track the object in small regions, the histogram may be sparse and not sufficient to represent the color distribution of the region. For instance, if the number of bins is set as $8 \times 8 \times 8$ and the region in image is 32×32, the expected number of pixels in each bin is only two, which is insufficient to represent the color distribution. To represent the color distribution, we model the histogram in color channel independently. Here, we select YC_bC_r as the color space, since the three channels are assumed independent. We divide the values in each channel into eight bins respectively. The expected number of pixels in each bin is 128, which can represent the color distribution more sufficiently. Another benefit of the modification is the computational efficiency when we compare the histograms between a particle and the target object, because the total number of bins is reduced to 24.

To represent the color histogram in several bins, another important task is how to map from a range of colors in the histogram to a bin. If the range is equally quantized for each bin and the histogram is compact, all pixels may fall into a small number of bins. In our cases, two different histograms cannot easily be distinguished. To cope with the problem, we first choose one histogram H as the reference one for histogram equalization. The equalization can be denoted as $z = M(H)$, where $M(.)$ is a function that equalizes the reference histogram H into an equalized histogram z, which is represented as a vector. The function $M(.)$ is then applied to another histogram H' to form a feature vector $z' = M(H')$. Based on the mapping, we can prevent the pixels from falling into the same bins for two slightly different color distributions.

Since the target object is moving, its appearance may change gradually. To adapt to the changes, the feature values should be updated for each frame as defined by

$$H_t = H_{t\text{-}1} \times 0.9 + Q_t \times 0.1, \tag{1}$$

where H_t and $H_{t\text{-}1}$ are the expected state observations at time t and $t - 1$, and Q_t is the histogram directly extracted from the estimated state at time t. The variables H and Q represent unequalized color histograms of the rectangles of both target object and estimated target region. Each of the two variables is defined as a 256×3-dimensional vector.

3.2 Adjustment of Number of Particles

The number of particles will greatly affect the search region of the target object in an image. In general, when more particles are selected, the tracker may become less efficient but more accurate. Therefore, we dynamically modify the number of particles to control the covering range in state space so that the state with the local maximum weight can be located. When we track a target with the particle filter, if the appearances of many regions are similar to the target object, the weights of particles will approximate a uniform distribution and have a higher entropy. If the target has been missed, the conditional probability $p(z_t \mid X_t = S_t^n)$ will be low and the distribution of weights will also approximate a uniform distribution. In these two cases, since we do not know where the target object is, a wide search window should be set. Therefore, a larger number of particles are needed. In another case, if there is only one region whose appearance is similar to the target object, the weights will concentrate on few particles and have a lower en-

tropy. In that case, a small search window is needed and a smaller number of particles is required.

To address the cases described above, we define the number of particles at time t (or the tth frame) based on the entropy as

$$N_t = C \cdot N_{t-1} \cdot \frac{-\sum_{n=1}^{N_{t-1}} \omega^{(n)} \log \omega^{(n)}}{-\log(1/N_{t-1})}, \tag{2}$$

where $\omega^{(n)}$ is the particle weight of the nth particle, and C is a constant to control the increase rate of the number of particles. For example, if C is set as two, the maximum number of particles at time t is $2 \cdot N_{t-1}$. In our experiments, the constant C is 1.2. To avoid the number of particles increasing or decreasing drastically, we limit the number N_t between (200, 1000) in our experiments.

4. THREE-PART HUMAN TRACKING AND CONSISTENCY CHECKING

To track a human reliably, the three parts, head, torso, and hip-leg, are tracked simultaneously. To design the tracking system, we first segment a person in a frame via an adaptive Gaussian background model [15], and then decompose the person into three parts. The positions of each part in the following frames are then predicted and updated using the particle filter described in the previous section. For each frame, after the positions of the three body parts are estimated by particle filters, consistency checking and adjustment of these body parts are performed to correct the abnormal body part. Finally, we perform an inter-person occlusion detection to avoid losing the target person when the person is occluded by other persons.

4.1 Human Part Extraction and Decomposition

In this study, we aim to separate the three parts with a high distinguishability. The distinguishability can be defined as the difference between the color histograms of two regions. We assume that the size ratios of the three body parts of most people are similar. As shown in Fig. 1 (a), we first locate two horizontal lines to separate the region of a person into three sections according to the predefined height ratio of the three parts, denote as H_h, H_t, and H_l. Then we move the two separation lines vertically to find the positions such that the three regions have the high differences in the mean colors. If a separation line moving up and down a short distance cannot achieve higher color difference than before, the moving is stop. If the separation line is moved to far away from the initial position, the separation line is reset to the initial position. Therefore, when the appearances of two body parts are much similar, this separation line may stop in the initial position. The foreground human region is accordingly separated by the two horizontal lines into three sections R_h, R_t, and R_l. The segmented foreground regions may include background regions or noise. Besides, the shapes of the three parts for different persons and different poses are varied. To achieve a higher reliability of the tracked parts, we will shrink the segmented regions according to the spatial distribution of the pixels in the three sections. A section R is shrunk into a smaller rectangle, called inner rectangle here-

(a) Initial horizontal separation lines of the person. (b) Results of final three parts of the person.

Fig. 1. An example of human parts decomposition of a person.

after, as the pixel set $\{(x, y) \mid C_x - S_x < x < C_x + S_x,\ C_y - S_y < y < C_y + S_y,\ (x, y) \in R\}$, where (C_x, C_y) is the center of the rectangle and (S_x, S_y) the covered range of the rectangle. The center (C_x, C_y) is defined as the mass center of all the foreground pixels in the section R, and the covering range (S_x, S_y) are the standard deviations of these foreground pixels in both x- and y-coordinates.

Fig. 1 (b) shows the three inner rectangles found for the position in Fig. 1 (a), in which the colors are more uniform. Note that the height of the inner rectangle of hip-leg is set to the half height of the segmented hip-leg, because the appearances of lower legs may vary significantly for different motions and dresses, which are not stable for tracking.

4.2 Tracking Failure Detection

The relative positions of the three body parts are limited in a certain range and the velocity of each part is also limited. Tracking failure will generate abnormal relative positions of estimated body parts, and the states will change irregularly in recent frames. If we can create a classifier to distinguish normally and abnormally tracked body parts, we can detect the event of tracking failure.

To detect the tracking failure, we also have to detect the failure component. In this study, we use support vector machines (SVM) [4] as classifiers to detect whether and which body part cannot be tracked properly. The SVM is a well-known classifier that finds a hyperplane in a higher dimensional space to separate data of two categories with the largest margin. To detect which part cannot be tracked, we design three SVMs for detecting the tracking failures of the three body parts. If the tracker fails to track two or three parts, the SVM failure detector for different body parts may become ineffective, since we cannot easily distinguish which part is abnormal by the relative positions. To cope with the problem, we design an additional SVM to determine whether the failure type is a single part failure or a multi-part failure.

The features used in an SVM are the estimated states of the three parts in the current frame and the relative state changes between the current frame at time t_0 and a previous frame at time $t_0 - \Delta t$. Here Δt is selected to make the state changes large enough (In our experiments, $\Delta t = 0.5$ seconds). The feature vector is defined as $[RS_H(t_0), RS_T(t_0), RS_L(t_0), RS_H(t_0) - RS_H(t_0 - \Delta t), RS_T(t_0) - RS_T(t_0 - \Delta t), RS_L(t_0) - RS_L(t_0 - \Delta t)]^T$, where the vectors $RS_H(t)$, $RS_T(t)$, and $RS_L(t)$ denote the relative state vectors of the three body parts in time t. The relative state vectors are defined as:

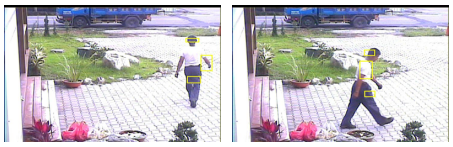
$$RS_{(K)}(t) = S_{(K)}(t) - \frac{1}{3} \sum_{I=H,T,L} S_{(I)}(t). \quad (K = H, T, L) \tag{3}$$

where $S_H(t)$, $S_T(t)$ and $S_L(t)$ are the estimated state vectors of head, torso and hip-leg parts.

To collect training samples, we apply particle filters to track the three body parts in several video sequences. We than manually label the training samples from these tracking results for each SVM. In our experiments, the number of training samples for each SVM is 150. The state vectors not covering the target body part are labeled as negative, while those falling inside are labeled as positive. The samples not satisfying these two criteria are eliminated; this ensures that the feature vectors of the two classes are distinguishable. In the tracker, since a misclassified tracking failure may cause error propagation and hard to be adjusted, we prefer a higher true-negative rate. Thus, we adjust the parameters of SVMs to achieve the goal.

4.3 Tracking Failure Adjustment

In case when a tracking failure is detected, we have to adjust the state of the target person. If two or three parts cannot be tracked, we will detect the foreground region around the previous tracked position of the target object, and re-initialize the tracking process. If the state of a single part is abnormal, we will use the other two body parts to adjust the position and size of the abnormal one. To keep the adjusted body part tracked in the following frames, the particle states and the appearance model (color histogram) must also be modified. If the abnormal body part still appears in the image, we can use the adjusted position and size to extract the particle states and the appearance model. However, if the failure is caused by occlusion, the appearances of the target person may not be correctly extracted as shown in Fig. 2 (b), and thus the system dynamic model should be used to track the person. In this case, the appearance model is not updated and the process of failure detection and adjustment is not performed either. The method of occlusion detection is discussed below.



(a) A prson with clothes colors similar to those of background regions.

(b) A person occluded by a pillar.

Fig. 2. Examples of two types of tracking failure.

4.3.1 Failure from inter-person occlusion

When two persons are both tracked, the occlusion event can easily be detected by checking whether the tracked body parts of the two persons are touching. If the answer is positive, we determine which one of them is occluded by measuring the similarity between the appearance of the body part and the observation model kept by the tracker.

4.3.2 Failure from background object occlusion

If a person is occluded by a fixed background object such as a pillar or a door, we

cannot detect the event since we only track the position of a person but not the background object. To cope with the problem, we can label manually the large and fixed background objects that may occlude moving humans. This is reasonable for a scene monitored by a fixed sensor.

The position of the tracking failure part can be adjusted according to the position of the other two parts. If the tracking failure part is the torso part as shown in Fig. 2 (a), we adjust the center of the torso to the middle of the other two parts and the size to the average of the other parts as follows:

$$\{X_T, Y_T, W_T, H_T\} = \frac{\{X_H, Y_H, W_H, H_H\} + \{X_L, Y_L, W_L, H_L\}}{2}. \quad (4)$$

Since the torso of a person is usually large enough, the adjusted rectangle usually lies inside the torso. Instead, the head of a person is usually smaller then the other two parts. Using similar adjustment method, its inner rectangle may contain background objects. If tracking failure happens in the head part, we will segment the foreground region $\{(x, y) | X_T - W_T \le x \le X_T + W_T, Y_T - 0.5 \cdot H_T \le y \le Y_T - 0.5 \cdot H_T + 2 \cdot (Y_T - Y_L)\}$, since this region covers most head regions in any poses. We then extract the inner rectangle from the foreground region by the method described in section 4.1. For the hip-leg part, its shape may have great variations. The adjustment method is similar to that of the head part, except that the foreground region is segmented below the torso part.

5. EXPERIMENTAL RESULTS

The proposed approach has been implemented on a personal computer. The experiments were performed on 22 video clips with 72 target persons in six scenes. The input images are in color with resolution of 720 × 480. The number of particles is set within 200 to 1000, which was automatically modified by using the entropy of particle weights.

The camera is mounted approximately three meters high and the angle between the camera and floor is smaller than 30°. The captured images of a person will be large enough to differentiate the body parts; the height and width of most persons are greater than 100 and 30 pixels, respectively.

5.1 Color Histogram Equalization

In this research, we propose an equalized observation model, as described in section 3.1. To test the effect of the model, we first define three rectangles on the head, torso, and hip-leg parts in 50 images as the target regions. We then compare the numbers of false-alarm target regions with and without equalization. A wrongly classified region is defined as the region that does not overlap with the target rectangle but its histogram feature is similar to that of the target one, whose similarity is less than a threshold. The similarity measurement between two histograms is defined as that in [2]. To define the threshold, we select the regions that are overlapped with the target one with more than half of size and then calculate the average histogram similarity. To calculate the number (or rate) of false alarms, we randomly define 1000 regions that do not overlap with the

target one as the test regions from each test image. Among the 50000 test regions (1000 regions × 50 images), 66 (0.13%), 79 (0.16%), and 45 (0.09%) regions were wrongly classified as belonging to a body part when we use the proposed histogram equalization; and 973 (1.95%), 734 (1.47%), and 725 (1.45%) without equalization, which demonstrates the effectiveness of our proposed equalization scheme.

5.2 Tracking Failure Adjustment

In our experiments on tracking the head, torso, hip-leg, and whole body on 20 video sequences longer than 100 frames, the tracking failure rates are 35%, 35%, 40%, 20% in the 100-th frame, respectively. Note that if the bounding rectangle of a target object contains less than a half of regions of other objects, the target object is regarded as tracked correctly; otherwise it is considered as tracked incorrectly. In the test results, most of the failures, about 100%, 100%, 87.5%, 75%, are propagated from previous frames. If we can keep the correctly tracked body parts and adjust the positions of other body parts, error will not propagate easily to the following frames and the accuracy can be improved.

0 40 200 220

(a) Tracking results with failure detection and adjustment.

0 40 200 220

(b) Tracking results without failure adjustment.

Fig. 3. The tracking results captured in an open space in front of a house.

Fig. 3 shows the human tracking results of a video clip captured in front of a house by applying our proposed method with failure adjustment and that without failure adjustment [2]. The numbers below the images denote the frame number after tracking initialization. Fig. 3 (a) is the tracking results using failure detection and adjustment, while Fig. 3 (b) is the results without failure detection. Since several non-human regions have similar appearance to parts of the target person, the regions may be mistaken as the body parts as shown in the 40th frame of Fig. 3 (b). In the frames, the appearances of the hip-leg part are similar to those of the stone. In the 200th and 220th frames, another background object is mistaken as the head part. In the 200th, and 220th frames, the tracking rectangles of the torso are slightly departed from the torso part, since the torso appearances are not similar to those in previous frames due to the motion of arms. Using our proposed method, we can detect the abnormal part and adjust its position as shown in Fig. 3 (a). Note that in the 200th frame, the head tracking rectangle is slightly departed from

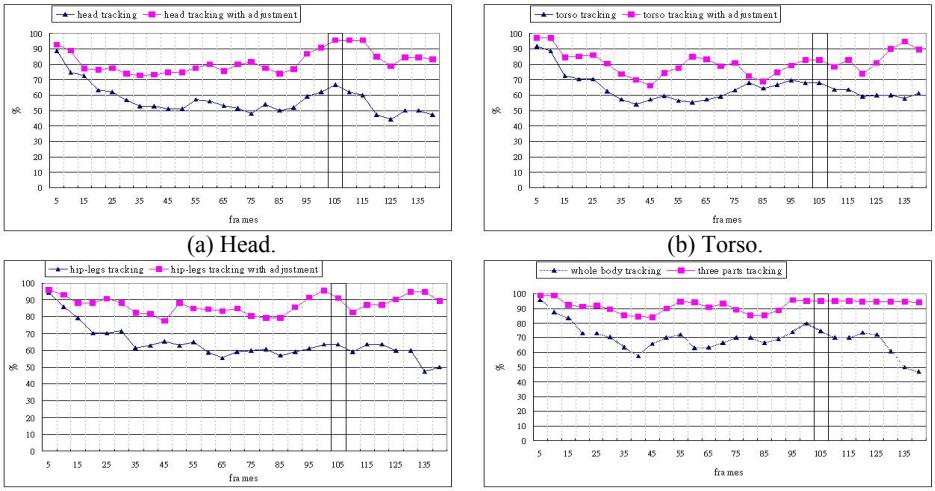
the head part. According to our failure adjustment, the head is corrected in the 220th frame.

5.3 Analysis of Tracking Accuracy

Figs. 6 (a)-(d) show the tracking accuracy curves of the three body parts and the whole body of the test video clips. Different persons appearing in a scene are recorded in different lengths of time intervals. Here, we test 72 sequences of target persons. The accuracy rate is defined as the ratio of the number of correctly tracked objects to the total number of detected objects. The error and correct rates are defined as those in section 5.2. The accuracy rate of the *i*th frame is the ratio of the persons tracked correctly from initialization to the frame. Since the sequence lengths of different persons are different, the denominators of the accuracy rates for the sequence of frames are varied.

The curves in Figs. 6 (a)-(c) show that the tracker with failure adjustment can improve the accuracy rate. The method of body part tracking without failure adjustment is similar to that proposed by Nummiar *et al.* [2] except for histogram extraction and weighting calculation, as described in section 3. In frame 105, for instance, the tracking accuracy rates with failure adjustment are 95%, 83%, and 91% for the three parts respectively. Note that the result curves are not monotonically decreasing, since the particle filter can adjust the target parts to the correct positions, and the numbers of frames that different persons walk in a scene are not the same. Comparing with the accuracy rates of the tracker without failure adjustment, 67%, 68%, and 64%, the tracking rates of the three body parts are improved about 28%, 15%, and 27%. In these samples, the torso parts of the target persons in the sequences longer than 70 frames are relatively stable than other parts as shown in Fig. 3. Therefore the accuracy curves of the two methods in the 80th frame of Fig. 4 (b) are much similar.

To test the effect of the multi-part tracker with failure adjustment for human tracking, we define the detected body region as the minimum bounding rectangle (MBR) that encloses the rectangles of the body parts. The single whole-body tracker is the same as



(a) Head. (b) Torso.

(c) Hip-leg. (d) Whole body.

Fig. 4. Tracking rates without and with failure detection.

that for the body part. Fig. 4 (d) shows the accuracy curves for the whole-body tracker (drawn as triangles) and the MBR of the whole body from the three-part tracker with failure adjustment (drawn as squares). In frame 105, for instance, the accuracy of the three-part tracker with failure adjustment is 95%. Comparing with the accuracy rate 75% of the whole-body tracker, the tracking rate is improved about 20%.

5.4 Multi-Person Tracking

When we track multiple humans, the main problem is occlusion. In the proposed method, we can use the system dynamic model to predict whether two persons are touching in a frame. In case of occlusion, we will find the occluded one and predict the target person until two persons are not touching anymore as described in section 4.3.1. Fig. 5 shows the images of a video that has multiple walking persons. In the first frame, the three persons from right to left are labeled as numbers 0, 1 and 2. In the fifth frame, person 1 hides the hip-leg part of person 2, and in the 12th and 14th frames, person 0 hides almost the whole body of person 2. In the images the white rectangles denote the tracked persons. The results show that our tracker can track target persons even during inter-person occlusion.

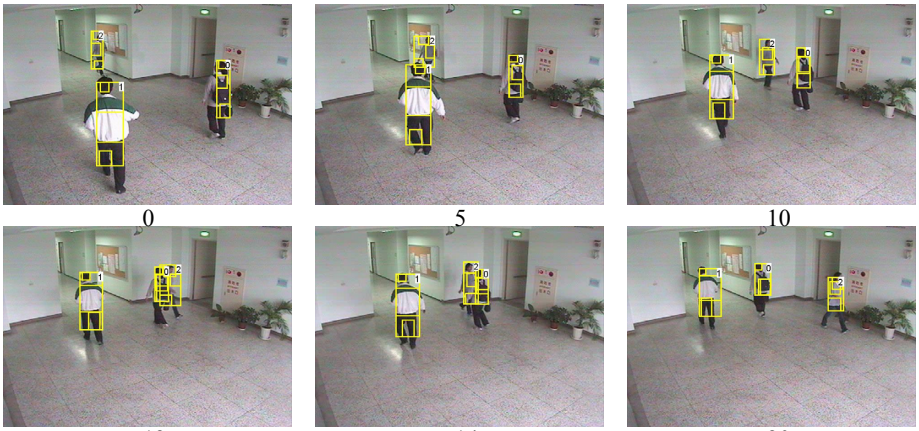


Fig. 5. The tracking results with failure adjustment and inter-person occlusion detection.

5.5 Tracking Failure Analysis

In our tracking system, we assume that the relative positions of body parts are fixed in a certain range. However, the assumption cannot be applied to track the body parts of a person when his posture is not trained, such as bending down as shown in Fig. 6 (a). In this situation, we can still track the same target person, but not his body parts correctly. Since these parts belong to the same person, we can still track the target person when he stands up.

In our failure adjustment, we use the background model to extract foreground regions. However, several false objects, such as shadow, may be regarded as foreground

(a) A person bending down.

(b) A sample affected by shadow.

Fig. 6. The tracking failure samples.

objects. In Fig. 6 (b), the target person is still tracked but the torso part is too large and the hip-leg part includes shadow. Usually, the false detected foreground regions only affect the tracking results several frames. When the three tracked body parts are not located in the correct relative positions, the failure adjustment scheme will adjust the positions of the body parts.

6. CONCLUSIONS

In this research, we have proposed a human tracking system. In the system, we decompose a human body into three parts, the head, torso, and hip-leg, and use color-based particle filters to track the three parts separately. We have also proposed an SVM-based method to detect the lost tracking part. In the tracking algorithm, we have used a particle filter for tracking an individual part. Since the particle number greatly affects the tracking performance and tracking speed, we use entropy of particle weights to modify the particle number dynamically. To further improve the tracking accuracy, we have designed a histo-gram equalization method for color histogram comparison. The experimental results show that the three parts tracking algorithm can improve the tracking accuracy significantly.

In this research, we assume that a human is standing up and the three body parts can be segmented from top to bottom. If a human crouches down or lies down, the body part decomposition may fail. Our experimental results show that in the case of failure, the three parts will not be labeled correctly. However, the failure will be adjusted after the target person stands up again, since we can detect the failure by SVM. To improve the body part decomposition, we may train detectors for different body parts. This is left for future research.

In our method, each of the three parts is tracked by a particle filter independently. The relative positions of the body parts are used to detect the tracking failure. We can reduce tracking failures by preventing the particles of abnormal poses to be generated. To achieve the goal, we needs to combine the state vectors of the three parts into a single vector to be tracked by a particle filter. Then the particle weights are adjusted according to the relative positions of the body parts. Also the behaviors of intruders defined on object appearances and the trajectories found will be analyzed. These are all left for future research.

REFERENCES

1. P. Perez, C. Hue, J. Vermaak, and M. Gangnet, “Color-based probabilistic tracking,”

in *Proceedings of European Conference on Computer Vision*, Vol. 1, 2002, pp. 661-675.
2. K. Nummiaro, E. Koller-Meier, and L. V. Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, Vol. 21, 2003, pp. 99-110.
3. A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 349-361.
4. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, Vol. 2, 1998, pp. 1-47.
5. T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, Vol. 104, 2006, pp. 90-126.
6. W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems*, *Man*, *and Cybernetics – Part C: Applications and Reviews*, Vol. 34, 2004, pp. 334-352.
7. A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, Vol. 38, 2006, pp. 1-45.
8. T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, Vol. 81, 2001, pp. 231-268.
9. D. M. Gavrila, "Visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, Vol. 73, 1999, pp. 82-98.
10. C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 747-757.
11. A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of IEEE*, Vol. 90, 2002, pp. 1151-1163.
12. A. Micilotta, E. Ong, and R. Bowden, "Detection and tracking of humans by probabilistic body part assembly," in *Proceedings of British Machine Vision Conference*, Vol. 1, 2005, pp. 14-20.
13. D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, 2003, pp. 564-577.
14. I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 809-830.
15. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, 1997, pp. 780-785.
16. S. L. Dockstader and N. S. Imennov, "Prediction for human motion tracking failures," *IEEE Transactions on Image Processing*, Vol. 15, 2006, pp. 411-421.
17. M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, Vol. 50, 2002, pp. 174-188.
18. A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance," in *Proceedings of Intelligent Vehicle Symposium*, 2004, pp. 1-6.

19. S. Ioffe and D. Forsyth, "Probabilistic methods for finding people," *International Journal of Computer Vision*, Vol. 43, 2001, pp. 45-68.
20. C. Chang, R. Ansari, and A. Khokhar, "Efficient tracking of cyclic human motion by component motion," *IEEE Signal Processing Letters*, Vol. 11, 2004, pp. 941-944.
21. B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, Vol. 75, 2007, pp. 247-266.
22. C. Lerdsudwichai, M. Mottaleb, and A. Ansari, "Tracking multiple people with recovery from partial and total occlusion," *Pattern Recognition*, Vol. 38, 2005, pp. 1059-1070.
23. S. Khan and M. Shah, "Tracking people in presence of occlusion," in *Proceedings of Asian Conference on Computer Vision*, 2000, pp. 263-266.

San-Lung Zhao (趙善隆) received the B.S., M.S., and Ph.D. degrees in Computer Science and Information Engineering form National Chiao Tung University, Hsinchu, Taiwan, in 1998, 2000, 2009, respectively. He is currently an Engineer in Industrial Technology Research Institute. His research interests include computer vision, image processing, and pattern recognition.

Hsi-Jian Lee (李錫堅) received the B.S., M.S., and Ph.D. degrees in Computer Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1976, 1980, and 1984, respectively.

From Aug. 1981 to July 2004, he had been with National Chiao Tung University as a Lecturer, Associate Professor and Professor. He was the Chairman of the Department of Computer Science and Information Engineering from Aug. 1991 to July 1997. From Jan. 1997 to July 1998, he was a Deputy Director of Microelectronic and Information Research Center (MIRC). From Aug. 1998 to July 2002, he was the Chief Secretary. Since Aug. 2004, he has been with Tzu-Chi University, Hualien. From Aug. 2004 to Feb. 2006, he was the Chairman of the Department of Medical Informatics. From Apr. 2006 to July 2010, he was the Dean of Academic Affairs. Since Feb. 2008, he has been the vice president of the university. He was the editor-in-chief of the International Journal of Computer Processing of Oriental Languages (CPOL) and associate editor of the International Journal of Pattern Recognition and Artificial Intelligence and Pattern Analysis and Applications. His current research interests include document analysis, optical character recognition, image processing, pattern recognition, digital library, medical image analysis, and artificial intelligence.

Dr. Lee was the president of the Chinese Language Computer Society (CLCS), Program Chair of the 1994 International Computer Symposium and the Fourth International Workshop on Frontiers in Handwriting Recognition (IWFHR) and was the General Chair of the Fourth Asia Conference of Computer Vision (ACCV), in January 2000.