International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 20, No. 3 (2006) 377–391

World Scientific
www.worldscientific.com

MULTIPLE HUMAN DETECTION AND TRACKING BASED ON WEIGHTED TEMPORAL TEXTURE FEATURES*

HEE-DEOK YANG†, SANG-WOONG LEE‡ and SEONG-WHAN LEE§

Center for Artificial Vision Research
Department of Computer Science and Engineering
Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Korea
†*hdyang@image.korea.ac*
‡*sangwlee@image.korea.ac.kr*
§*swlee@image.korea.ac.kr*

In this paper, we present a method of tracking and identifying persons in video images taken by a fixed camera situated at an entrance. In video sequences a person may be totally or partially occluded in a scene for some period of time. The proposed approach uses the appearance model for the identification of persons and the weighted temporal texture features. The weight is related to the size, duration as well as the number of persons adjacent to the target person. Most systems have built an appearance model for each person to solve occlusion problems. The appearance model contains certain information on the target person. We have compared the proposed method with other related methods using color and shape features, and analyzed the features' stability. Experimental results with various real video data sequences revealed that real time person tracking and recognition is possible with increased stability in video surveillance applications even under situations of occasional occlusion.

Keywords: Multiple object tracking; video surveillance; multiple people detection; appearance model; temporal texture; human activity recognition.

1. Introduction

Tracking humans in video sequences is an important task in many applications such as video surveillance and virtual reality interfaces, since they are the primary actors in these task domains. In the past few years, a number of real time systems have been developed to detect and track people. VSAM[2] is a CMU development used for tracking people and moving objects. The system's infrastructure involves fourteen cameras. Haritaoglu *et al.*[5] introduced the system W4 for detecting and

*A preliminary version of this paper was presented at the 17th International Conference on Pattern Recognition, Cambridge, United Kingdom in August 2004. This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.
§Author for correspondence.

tracking multiple people or part of their bodies. Darrell *et al.*[3] used disparity and color information to extract and track individual persons.

More recently, there has been increasing interest in integrating the temporal information contained in a tracking system.[1] Among the systems listed above, W4[5] system already employs temporal information to improve the performance of tracking and identification. With temporal information, the search range can be narrowed down significantly. In general, tracking people consists of two subtasks, viz. target detection and verification.[1,8] Also, a means is required to answer the questions: "Has the person been detected correctly?" and "Do the people detected in the previous frame match any of the people detected in the current frame, or vice versa?" Usually the number of candidate persons in a scene is small, but it is not rare to process a large number of candidates for an equivalent number of candidates in the previous frame.

The task of human tracking is complex if there are many people included in a scene, and even more so if they are interacting with each another. Figure 1 shows the traditional tacking system. Traditional tracking tasks can be classified into three types according to the time scale, namely short-term, medium-term and long-term.[7]

The *short-term tracking* applies when the target stays within the scene. In this type of tracking, we estimate the second order motion for the purpose of tracking the target.

The *medium-term tracking* applies if the target human becomes occluded or if they re-enter the scene within a few minutes. The appearance model is essential to accomplish medium-term tracking. The appearance model includes many features used to represent a human being, such as his or her shape, texture, intensity, color, and face pattern.

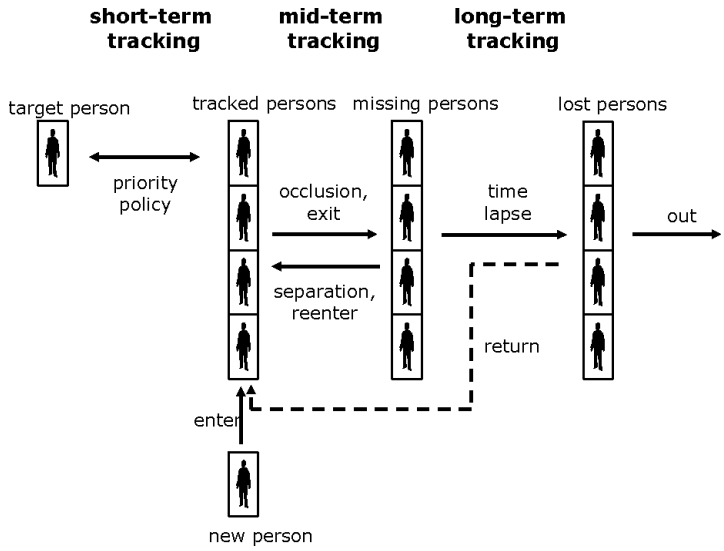


Fig. 1. Traditional tracking system.

The *long-term tracking* is an extension of medium-term tracking. In this case, the temporal scale is extended to hours or even days. Most of the features used in medium-term tracking are unsuitable for this type of tracking, because they become unstable on such a large temporal scale. The only stable, and therefore useful, feature is the facial pattern.

In this paper, we propose a person tracking method based on the appearance model using the temporal texture features of the image sequence. The temporal texture is a set of pairs of a texture value and an associated weight. The weight connotes or summarizes the size, duration, position, velocity and frequency of appearance of the texture region, as well as the number of people adjacent to the target person. Finally, we employ a simple method of recognition for the verification task.

2. Overview

In this paper, we propose a novel method of detecting and tracking people simultaneously in a surveillance environment. The overall organization of the system is shown in Fig. 2. The system consists of three parts: human detection, human tracking and face recognition. The human detection phase, in turn, is decomposed into several sequential subtasks consisting of motion detection, candidate region detection, and human detection. In human tracking, the human hypotheses are confirmed for identification, and then passed to match with human tracked in the preceding frames. Finally, the face recognition task is achieved by a simple method which consists of face detection and identification. Note that there are feedback loops in the candidate region detection and human tracking modules.

3. Human Detection

The goal of human detection is to locate and isolate all of the persons in a scene. This task is divided into two parts: the detection of candidate regions and the detection

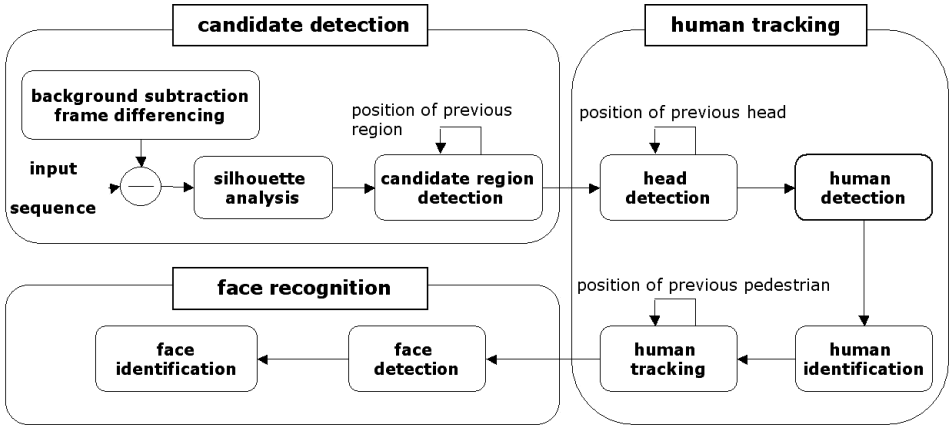


Fig. 2. System overview.

of individual persons. A candidate region may contain one or many moving persons, and the human detection task explicitly divides the region into individual persons.

3.1. *Candidate region detection*

Candidate region detection is a simple process whose purpose is to locate any moving objects in a scene. Here, we utilize the methods of adaptive background subtraction and three-frame differencing.

Let $I_n(x)$ represent the intensity value at pixel position x and time n, $B_n(x)$ denote the current background intensity value at pixel x and time n learned by observation over time, and $T_n(x)$ the difference threshold. $B_0(x)$ is initially set to the value of the first frame, i.e. $B_0(x) = I_0(x)$, and $T_0(x)$ is supplied externally. $B_n(x)$ and $T_n(x)$ are updated over time as follows[2]:

$$B_{n+1}(x) = \begin{cases} \alpha B_n(x) + (1-\alpha) I_n(x), & x \text{ is nonmoving} \\ B_n(x), & x \text{ is moving} \end{cases} \tag{1}$$

$$T_{n+1}(x) = \begin{cases} \alpha T_n(x) + (1-\alpha)(5 \times |I_n(x) - B_n(x)|), & x \text{ is nonmoving} \\ T_n(x), & x \text{ is moving} \end{cases} \tag{2}$$

where α is a time constant denoting the relative magnitude of the influence from the new input.

The three-frame difference is used to detect moving pixels. A pixel x is moving if it satisfies Eq. (3).

$$(|I_n(x) - I_{n-1}(x)| > T_n(x)) \quad \text{and} \quad (|I_n(x) - I_{n-2}(x)| > T_n(x)). \tag{3}$$

3.2. *Human segmentation*

Usually, a foreground region contains more than one person, in which case we need to partition it into subregions containing individual persons. In the first step of human detection, we apply morphological operators such as erosion and dilation to eliminate noise. The resulting image is referred to as a silhouette image.

The subsequent segmentation step mainly concerns the process of locating heads in the silhouette image, by shape analysis and vertical projection. By observing the curvature of the boundary of a silhouette, the part of a person that looks like a head may be located. The point of maximal convex value is a good indicator of the head, while the point of minimal concave value is similarly a good indicator of the separation between two persons.

The system divides the region into subregions, each corresponding to an individual person whose head was detected in the preceding stage. Finally, the torso axis is estimated to be the vertical line descending from the center of the head, and the segment containing each person is estimated according to the distance of the pixels from the torso axis.

The detected region must be further segmented into individual persons, if there are more than one head in the region. Since the only information known to us at

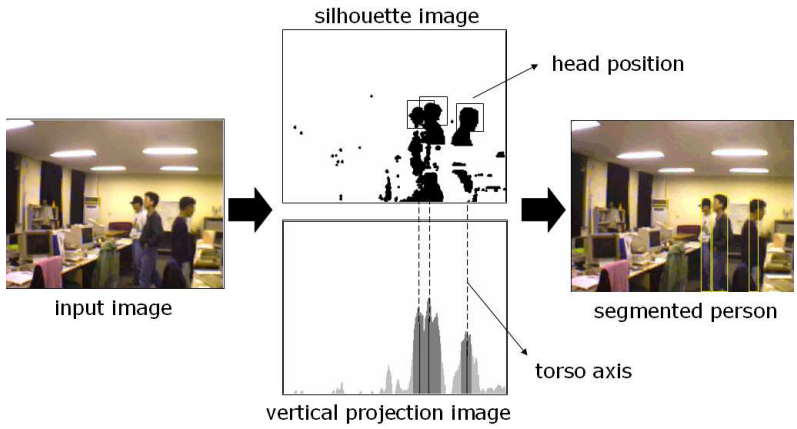


Fig. 3. Process of human detection.

this stage is the head positions, the segmentation task begins with the heads. The torso axis is estimated from the position of the head, and the distance from the axis to each pixel in the detected region is computed. The distance designates, for each pixel, which axis is nearest to the pixel. The distance used here is described in Ref. 6 as follows:

$$n(x,y) = \frac{\alpha(x,y)}{\sum_z(\alpha(x,z))} \quad \text{where } \alpha(x,y) = \frac{\min_z(d(x,z))}{d(x,y)}. \tag{4}$$

In Eq. (4), $n(x,y)$ is the normalized distance value of the path distance $d(x,y)$.

Figure 3 shows an example of human segmentation.

The person-segmentation and person-identification modules are bounded with one box in Fig. 2, which means that the person-identification module is called upon only when the person segmentation module is involved in the detection process. More specifically, the person segmentation module is triggered when there are more than two heads in a group of people.

4. Human Tracking

The temporal information along the video frame sequence is conjectured to have a potential for greater accuracy and speed-up. This point will be discussed here together with a solution for brief occlusions.

4.1. *Temporal texture feature*

For accurate and faster person tracking, we selected the temporal texture feature as a supplement to the appearance model. This should capture the inter-frame correlation and continuation for successful tracking. Let us start with the definition; the temporal texture T is defined as a set of pairs of texture values t and its temporal

weights w as

$$T = \{(t_1, w_1), (t_2, w_2), \ldots, (t_i, w_i)\} \tag{5}$$

where i is the number of intensity clusters.

The texture value t measured in intensity units, represents the mean or center of a texture cluster in the intensity space. The first step in calculating the temporal texture is the clustering of textures. For each person, an intensity histogram is created and the mean intensity value of a person is calculated. In this way, two or three intensity clusters can be obtained for each person.

The temporal weight w is the coherency of the intensity cluster. It is a function of the size of intensity cluster and variance of the intensity values. The relation between each parameter is

$$w_i \propto \frac{1}{\Delta s_{t_i} \cdot \Delta m_{t_i}} \tag{6}$$

where Δs_{t_i} is the difference in the number of pixels at texture value t_i and Δm_{t_i} is the difference in the mean number of pixels at texture value t_i.

The velocity of the person is temporal in that it changes in every frame, while the size of the person is presumably not temporal and, hence, not explicitly used in the model. This has a nontrivial influence on the amount of computation. For this reason, we include it as a temporal texture feature.

The weight W is the coherency of the intensity cluster. It is a function of the size of the intensity clusters and the variance of the intensity values. The relationship between each parameter is

$$W_n \propto \frac{\Sigma w_n \cdot TA_n}{\Gamma(n) \cdot \Delta S_n \cdot \Delta p_n \cdot \Delta v_n} \tag{7}$$

where n is the number of frame in the image sequence, S_n is the number of pixels in the region containing the person, $\Delta S_n = S_n - S_{n-1}$, p is the position of the person and $\Delta p_n = p_n - p_{n-1}$, $v = \Delta p_n / \Delta t$, $\Delta v_n = v_n - v_{n-1}$, and finally Δt is the time interval between successive frames.

The adjacency function Γ measures the 2D geometric adjacency or separation between two persons. It is related to the shape of the target person and the distance from the other person, as:

$$\Gamma_p(t) = \sqrt{\frac{A_p}{\left(x_p^t - x_q^t\right)^2 + \left(y_p^t - y_q^t\right)^2}} \tag{8}$$

where x_p, y_p is the center coordinate of the target person at time t, and A is the length of the target in the direction that the adjacency occurs.

4.1.1. *Texture feature used in temporal texture model*

The intensity clustering algorithm and intensity distance are related to the accuracy and efficiency of the texture feature method. There are many intensity clustering

algorithms. In order to choose the most appropriate algorithm for the proposed method, the following considerations must be taken into account:

- the intensity clustering algorithm must be as inexpensive as possible in terms of computing time;
- the intensity similarity measure must be robust to illumination and noise

In the proposed method, an intensity cluster is selected by histogram analysis. Each pixel that is found to be part of the foreground region is represented as an intensity value.

To construct histogram bins, the intensity value is divided equally and the histogram bin whose value exceeds the threshold is selected as the representative intensity cluster. The average intensity value of all pixels in a selected histogram bin is the intensity feature of the proposed method. When dividing histogram bins of the same size, an error can occur where one cluster is divided into two clusters. To solve this problem, all of the intensity values of a selected cluster are compared with each other and the two closest intensity clusters are merged. Figure 4 shows the process of the proposed tracking algorithm.

4.1.2. *Tracking framework with temporal texture*

The tracking framework for the appearance model with temporal textures is described in this section. In the short-term tracking process, second order motion estimation is used for the continuous tracking of an object. Motion estimation and matching with texture features are used simultaneously to find any correspondences between two consecutive frames when the target is occluded or out of sight. The positional similarity determined by motion estimation and the texture similarity determined by means of the temporal texture features are added with different weights according to the result of the adjacency test.

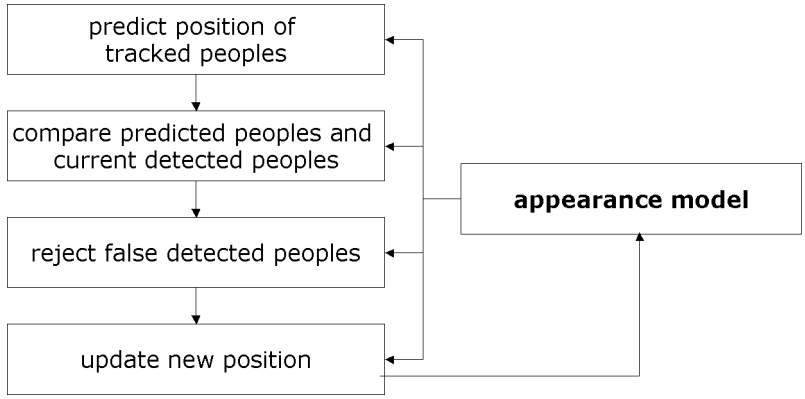


Fig. 4. Procedure of tracking framework.

- **Case 1 — isolated**: only the positional information is used with the second order motion model.
- **Case 2 — partially occluded**: the positional information and texture features are used simultaneously.
- **Case 3 — totally occluded**: the occluding person is tracked as if he or she were an isolated person, and all references to the occluded person are replaced by those of the occluding person.
- **Case 4 — after occlusion**: the positional information and texture features are used; the positional information is used for the occluding person.
- **Case 5 — after missing**: the texture features and positional information are used, that is, the position at which he/she was missing.

The occluding person who is used in Cases 3 and 4 plays an important role in mid-term tracking. In mid-term tracking, the positional information of the occluded person is not continuous. Since the position of the occluded person is normally the same as that of the occluding person in the case of total occlusion, the positional information of the occluding person can be considered to be the same as that of the occluded person.

4.2. *Occlusion detection*

Let us consider the situation in which two persons are being tracked through occasional occlusions. We can determine an approximate depth for each target with a separate model for occlusion such as:

$$O_{AB} = \max \sum_{x \in R} \delta(I_{AB}(x), D_{AB}(x)) \tag{9}$$

where $\delta(a, b) = \begin{cases} 0, & a = b \\ 1, & \text{otherwise} \end{cases}$, A and B represent two different persons, R represents the overlapping region between the two persons, and $D_{AB}(x)$ denotes the nearer person

$$\arg \min_{\{A,B\}} \{|x - Ac|, |x - Bc|\}$$

where Ac and Bc are the center of the two people, $I_{AB}(x)$ denotes the person whose intensity at x, i.e. $I_A(x)$ or $I_B(x)$, is closer to the current observation $I(x)$

$$\arg \min_{\{A,B\}} \{|I(x) - I_A(x)|, |I(x) - I_B(x)|\}.$$

Figure 5 shows an example of occlusion.

5. Experimental Results and Analysis

5.1. *Experimental environment*

Our tracking system was implemented and evaluated on a Pentium IV-1.7 GHz PC with Microsoft Windows 2000. The video images were acquired at a rate of 30 fps using a Meteor II frame grabber and Jai 3CCD camera with the resolution set to

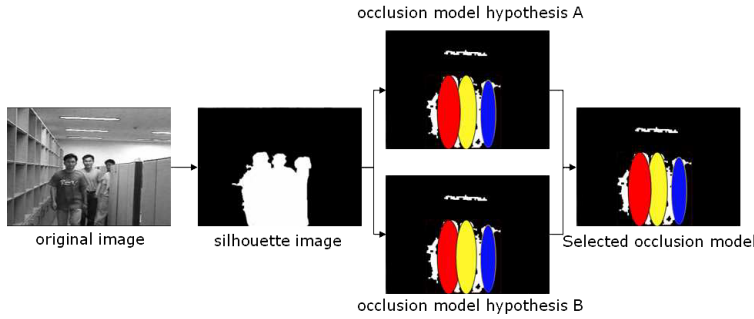


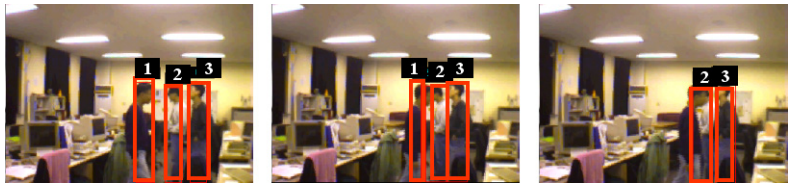
Fig. 5. An example of occlusion model.

320×240 pixels. The video captured two or three persons entering an office room together. The data set includes 600 frames in total.

5.2. *Experimental results*

Figure 6 shows some of the results obtained for the tracking of three people in different occlusion situations. It can be seen that two persons are continuously tracked throughout the occlusion. The last frame in Fig. 6(a) shows the case in which one person is missing because two people overlap each other. In this case, the heads of the two people are overlapped, and one head is lost. Figure 6(b) shows the result obtained in the same situation by incorporating temporal weights.

Table 1 shows a brief description about the content of the test data and the results of the process of detection and tracking. The number of persons is the number of individual persons who appeared in the scene at any one time. The numbers of false detections and false trackings are the numbers of error frames in which one



(a) Result of detected person without temporal weight

(b) Result of detected person with temporal weight

Fig. 6. Tracking examples (scene I), without top and with bottom temporal weights.

Table 1. Tracking result from three sample scenes; the detection error includes both person misses and ghosts (false positives), and the tracking error includes loss or incorrect location of persons, counted in frames from the total frames.

	Number of Persons	Sum of Frames	Detection Errors	Tracking Errors
scene I	3	200	2	1
scene II	2	150	4	2
scene III	3	250	8	3

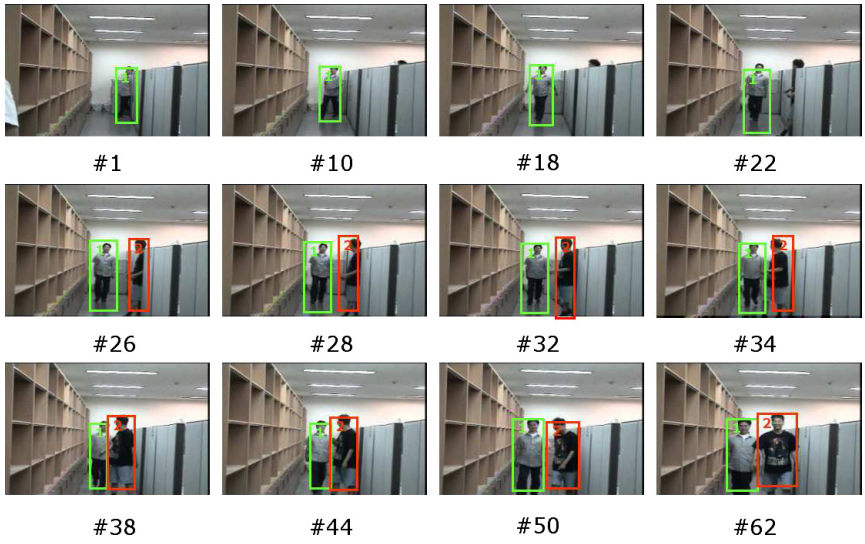


Fig. 7. Tracking examples (Scene II).

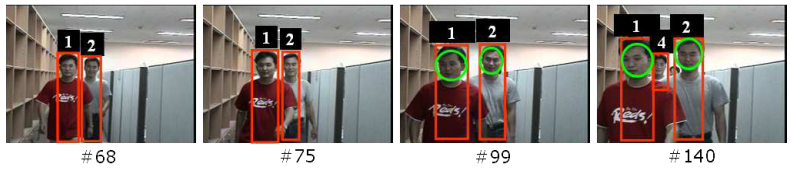
or more persons are missing or incorrect, respectively. In the case of false tracking, the error may affect the next frame. In the current experiment, we simply ignored all the consequential errors.

Figure 7 shows a small sample sequence of shots containing two persons with simple movement. Figure 8 shows another example containing four persons with very complex patterns of movement. The frame rate of the data used in Fig. 7 was 30 frames per second. However, our system could not handle such a high frame rate.

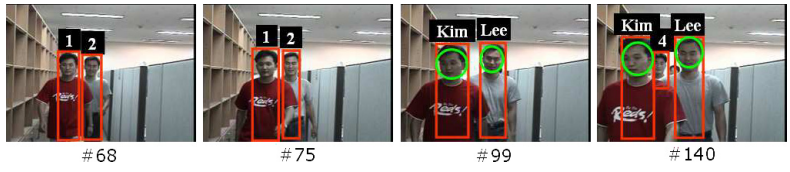
Now let us assume that we have a complete and high-performance tracking system. In order to evaluate the feasibility of developing an intelligent integrated system, we combined the tracking system with our existing face recognition system for the purpose of creating a real time surveillance task. The face recognizer is based on the support vector machine described in Ref. 10. For a detailed description, please refer to this paper.

In the current setup of the system, we have not yet carried out a systematic evaluation. Thus, for now, the best result we can provide is a few snapshots of the working system. Figure 9 shows some frames containing the results

Fig. 8. Tracking examples (Scene III).



(a) Result of human tracking without face recognition



(b) Result of human tracking with face recognition

Fig. 9. Example of person tracking and face recognition.

obtained from the process of person tracking and, if the person is sufficiently near, identification.

Note that the reference to person #1, in frame #68 is changed to "Kim" in frame #99, and that person #2 is changed to "Lee". The information about the name and the face is stored in the database.

5.3. *Performance analysis*

In the final set of tests, we compared the performance of the texture feature used in our system with the color feature used in Ref. 5 and the shape feature of the temporal templates used in Ref. 5. The quantitative comparison of the performance is based on the following measure described in Ref. 7 as follows:

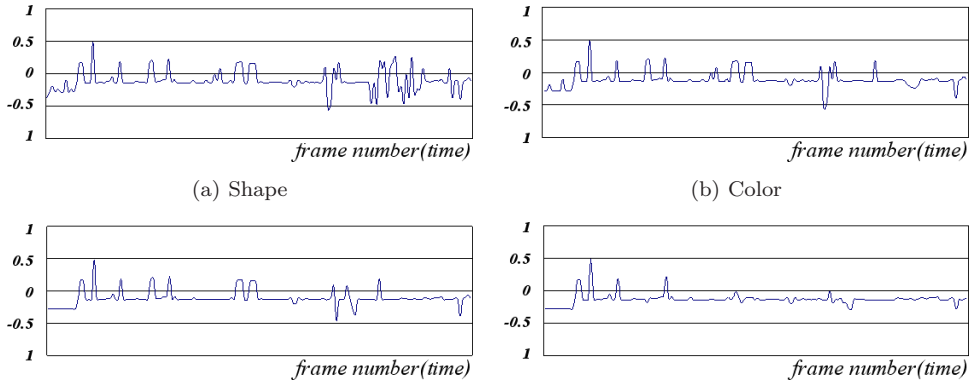
$$\frac{\sum_{x,y\in P}(M_P - I(x,y))}{|P|} \tag{10}$$

where P is a set of foreground pixels of the target person, $M_P = \frac{\sum_{x,y\in P} I(x,y)}{|P|}$ is the mean intensity in P, $I(x,y)$ is the pixel intensity for the texture feature at (x,y), (the color value for the color feature, 0 or 1 for the shape feature), and $|P|$ is the number of pixels of the target person in P.

Figure 10 shows the variation of each feature as a function of time for person #1 in Fig. 6. The horizontal axis is derived from the discrete time frame, while the vertical axis is derived from the normalized variance measured by Eq. (10). Figure 10 shows that the texture feature is more stable than the shape or color feature. Figure 10(d) implies that the stability can be improved if we introduce a face recognizer and an occlusion model.

Table 2 shows the time required for each feature involved in the tracking system. The average processing time was calculated using 600 frames.

Figure 11 shows the evaluation using the occlusion model. Figure 11 was the result of the experiments performed with person "2" of Fig. 6. Figure 11(a) shows



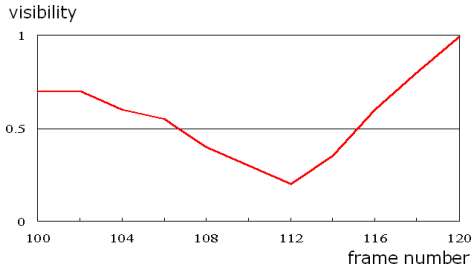
(a) Shape (b) Color

(c) Texture (d) Texture + face pattern

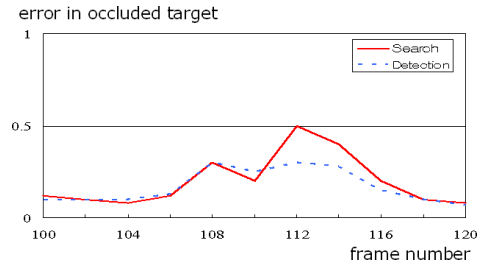
Fig. 10. Comparison of the stability of each feature.

Table 2. Average processing time with sample scenes, 600 frames.

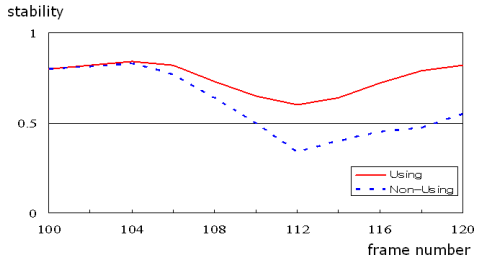
	Shape	Color	Texture	Texture + Face Pattern
Average processing time (sec)	0.11	0.1	0.08	0.25



(a) Visibility of occluded target



(b) Error in occluded target



(c) Stability

Fig. 11. Stability of occulsion model.

the visibility of the person. Figure 11(c) shows the stability of the person. In order to evaluate the performance of the proposed method, we analyzed the regions in which the persons were detected, before and after the application of the occlusion model. We measured the stability of the defined object for each region found to contain a person as the number of pixels correctly detected/number of pixels detected for each region by means of Eq. (11) described in Refs. 4 and 11. The result shows that using the occlusion model provides more stable results than not using it.

$$I_t = \frac{I_{t-1} * (t-1) + C(t)}{t} \tag{11}$$

where $C(t)$ is the intensity value at time t, which is equal to the value used in Eq. (10).

6. Conclusion and Further Research

In this paper, we proposed a new human detection and tacking algorithm based on an appearance model using temporal textures in video streams. The texture features include the temporal texture value, size, position, velocity, and frequency of the texture value region, and they are temporal in that they vary over time with associated weights and an occlusion model.

The experimental results show the stability of the proposed method of tracking. The temporal features are not complete. Rather they can be further generalized with any kind of features in the appearance model for the person tracking process. One problem with the use of the temporal texture arises when the people

are in uniform or dressed in clothes with the same color and intensity. In order to solve this problem, we introduced other features such as the position and face recognition result. Although face recognition is the most discriminating feature, the computational cost of face analysis is too expensive and the accuracy is often not sufficiently high. However, we applied face recognition to our system when the region of the person is sufficiently large for face recognition to be applicable. By applying face recognition, we could increase the discrimination and the security level.

Future research will be focused on the development of a tracking system using multiple cameras. By using multiple cameras, we believe that we can reduce the tracking error and expand the tracking area.

References

1. L. Baoxin and R. Chellappa, A generic approach to simultaneous tracking and verification in video, *IEEE Trans. Imag. Process.* **11**(5) (2002) 530–544.
2. T. Collins *et al.*, A system for video surveillance and monitoring:VSAM finial report, Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University (May 2000).
3. T. Darrell *et al.*, Integrated person tracking using stereo, color, and pattern detection, *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, Santa Barbera, CA (1998), pp. 601–608.
4. A. Elgammal and L. S. Davis, Probabilistic framework for segmenting people under occlusion, *Proc. 8th Int. Conf. Computer Vision*, Vancouver, Canada (July 2001), pp. 145–152.
5. I. Haritaoglu, D. Harwood and L. S. Davis, W4: Who? When? Where? What? A real time system for detecting and tracking people, *Proc. 3rd Int. Conf. Face and Gesture Recognition*, Nara, Japan (April 1998), pp. 222–227.
6. I. Haritaoglu, D. Harwood and L. S. Davis, Hydra: Multiple people detection and tracking using silhouettes, *Proc. 2nd IEEE Workshop on Visual Surveillance*, Fort Collins, Colorado (June 1999), pp. 6–13.
7. H.-K. Roh and S.-W. Lee, Multiple people tracking using an appearance model based on temporal color, *Proc. 1st Int. Workshop on Biologically Motivated Computer Vision*, Seoul, Korea (May 2000), pp. 369–378.
8. S. S. Intille, J. W. Davis and A. F. Bobick, Real-time closed-world tracking, *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Puerto Rico (Junuary 1997), pp. 601–608.
9. C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, Pfinder: real-time tracking of the human body, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(7) (1997) 780–785.
10. D. Xi and S.-W. Lee, Face detection and facial feature extraction using support vector machines, *Proc. 16th Int Conf. Pattern Recognition*, Quebec City, Canada (August 2002), pp. 209–212.

Hee-Deok Yang received his B.S. degree in computer science from Chungnam National University, Daejeon, Korea, in 1998, and his M.S. degree in computer science and engineering from Korea University, Seoul, in 2003. Currently, he is a Ph.D. degree candidate in the Department of Computer Science and Engineering, Korea University.

His research interests include motion and object detection, object tracking and gesture recognition.

Sang-Woong Lee received his B.S. degree in electronics and computer engineering from Korea University, Seoul, in 1996, and his M.S. degrees in computer science and engineering from Korea University, Seoul, in 2001. Currently, he is a Ph.D. degree candidate in the Department of Computer Science and Engineering, Korea University, Korea.

His research interests include face recognition, robot vision and pattern recognition related fields.

Seong-Whan Lee received his B.S. degree in computer science and statistics from Seoul National University, Korea, in 1984; and M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology in 1986 and 1989, respectively.

From February 1989 to February 1995, he was an Assistant Professor in the Department of Computer Science at Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Seoul, and now he is a full professor. Prof. Lee is also a visiting professor at the Artificial Intelligence Laboratory of MIT.

Prof. Lee was the winner of the Annual Best Paper Award of the Korea Information Science Society in 1986. He obtained the Outstanding Young Researcher Paper Award at the 2nd International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Professor Award from Chungbuk National University in 1994. He obtained the Outstanding Research Award from the Korea Information Science Society in 1996. He also received an Honorable Mention of the Annual Pattern Recognition Society Award for an outstanding contribution to the *Pattern Recognition Journal* in 1998.

He is a fellow of International Association for Pattern Recognition, a senior member of the IEEE Computer Society and a life member of the Korea Information Science Society.

He has more than 200 publications on computer vision and pattern recognition in international journals and conference proceedings, and has authored 10 books.

His research interests include computer vision, pattern recognition and neural networks.