Active Object Tracking using Image Mosaic Background

Young-Kee Jung, *Member, KIMICS,* **Dong-Min Woo,** *Nonmember*

Abstract—In this paper, we propose a panorama-based object tracking scheme for wide-view surveillance systems that can detect and track moving objects with a pan-tilt camera. A dynamic mosaic of the background is progressively integrated in a single image using the camera motion information. For the camera motion estimation, we calculate affine motion parameters for each frame sequentially with respect to its previous frame. The camera motion is robustly estimated on the background by discriminating between background and foreground regions. The modified block-based motion estimation is used to separate the background region. Each moving object is segmented by image subtraction from the mosaic background. The proposed tracking system has demonstrated good performance for several test video sequences.

Index Terms—Active Object Tracking, Image Mosaic, Camera Motion Estimation, Object Detection

I. INTRODUCTION

For the automated surveillance system, we use a camera to watch moving objects in the restricted area. If objects move outside the field of view, the camera should pan or tilt such that they always stay within its filed of view. In those applications, motion detection and tracking for moving objects play quite important roles.

The moving object detection and tracking can be applied to the popular video conferencing environments. In general multipoint video conferencing environments, each participant has its own background, and the conferencing environment looks not concordant at all. This kind of video conferencing environment is quite different from traditional conference. To overcome this disadvantage, we can create a virtual environment and put the segmented objects in it, so the object-based videoconference will look more realistic. The technique can be also applied to surveillance applications to detect and segment out the intruding objects in the home of office.

Although various research works have addressed these application areas [1-5], it is difficult to design general and robust solutions to the problems involved. This difficulty mainly stems from the complicated relationship between the motion of objects in the 3-D scene and the apparent motion of brightness patterns in the sequence of 2-D projections of the scene. Information about the relative depth of objects is lost in the projection, and the observed motion in the image plane can result from other phenomena than the object motion in the scene, such as changes in the lighting conditions.

Moreover, the presence of observation in the 2-D image sequence is in itself a non-trivial task because of the presence of observation noise, occlusions and temporal aliasing. Especially, for the active camera, because the moving camera creates image changes due to its own motion, object tracking with the mobile camera is a very challenging task.

In this paper, we attempt to address these problems. First, we trace a moving object based on the image mosaic background for wide surveillance. Second, we utilize the affine model to generate the image mosaic background. The image mosaic is a panoramic image that is constructed from multiple frames in the video sequence [6,7]. The affine model provides greater flexibility in modeling the global motion, being able to represent rotation, dilation and shear as well as translation. Third, the camera motion is robustly estimated on the background by discriminating between background and foreground regions. Therefore, the camera motion estimate is not spoilt by the presence of outliers due to foreground objects whose motion is not representative of the camera motion.

II. PROPOSED TRACKING ALGORITHM

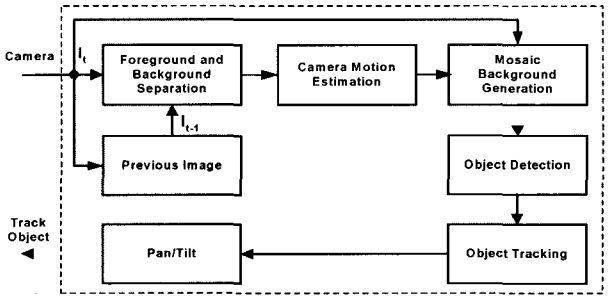


Fig. 1 Proposed object tracking algorithm.

As shown in Fig. 1, the proposed tracking system consists of five functional parts: foreground and background region separation, camera motion estimation, mosaic background, object detection & tracking, and control of the pan-tilt camera. The system first identifies background and foreground regions based on dominant motion estimates. Camera motion is then estimated on the background by

Manuscript received January 2, 2004.

This work was supported by grant No. R01-2002-000-00336-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

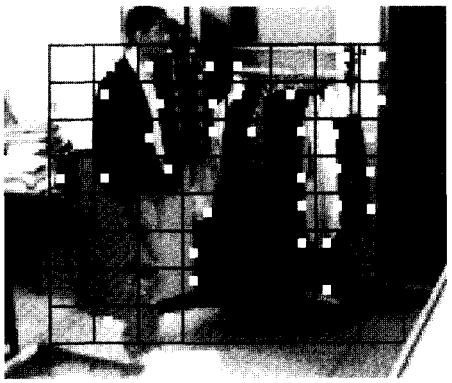
Y. K. Jung is with the Department of Computer Engineering, Honam University, Kwangju, Korea (e-mail: ykjung@honam.ac.kr)

D. M. Woo is with the Department of Information Engineering, Meongji University, Yongin, Korea (e-mail: dmwoo@mju.ac.kr)

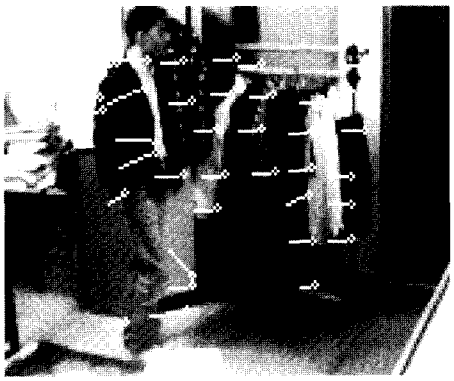
applying the parametric affine motion estimation algorithm. The image mosaic background content is integrated in a single image. Finally, after we detect the moving object, we trace it at the center of the camera.

4. *Background and Foreground Separation*

Discrimination between background and foreground is based on block-based motion estimation. A dominant motion is extracted by a clustering of the block vectors. Then, regions moving according to the dominant motion are identified as background, and otherwise as foreground. This separation has the following two steps: block-based motion estimation and background region extraction.



(a)



(b)

Fig. 2 Block-based Motion Estimation (a) the selected 9×9 region for block motion estimation (b) the extracted block vectors

1) Block-based Motion Estimation

In this paper, the modified block-based estimator is used to track changes of individual blocks while the global motion estimation step is introduced for deriving a single representative affine motion. Each frame of 320×240 resolution is divided into multiple 32×24 blocks, and for block motion estimation, a 9×9 window region with maximum standard deviation is extracted within each block, as shown in Fig. 2.

However, in low contrast areas, resulting motion vectors are unreliable. In order to overcome this problem, we apply the activity criterion to filter out unreliable blocks with lower standard deviation than a certain threshold value. The 9×9 template that was extracted is correlated in the search region. After we locate the correlation peak, a motion vector is associated with each block. The block motion vector holds the displacement of the block between the current and the previous frames.

2) Background Region Extraction

In order to extract the background motion, we compute a dominant by the following steps:

(a) For all block motion vectors, count the number of times that a motion vector is used.

(b) Obtain the most and second-most used motion vectors.

(c) Average the two candidates motion vectors.

Finally, if the motion of the block is similar to the dominant motion, we will consider this block as the background block; otherwise, blocks of the foreground block or noise block are removed

B. Camera Motion Estimation

After discriminating the background motion from other motions, we estimate the camera motion in the background. In this way, the camera motion estimate cannot be spoilt by the presence of outliers due to foreground objects, whose motion is not representative of the camera motion.

The camera motion can be modeled by a parametric affine motion model with six parameters. We first estimate the six parameters using the least square method from the background motion vectors. Once motion parameters are obtained, we compensate the camera motion through the inverse affine motion transformation.

Let (x, y) be a block vector position in the previous frame, and (x', y') be the position in the current frame. Then, we can represent the motion vector (v_X, v_Y) by

$$\begin{pmatrix} v_X(x,y) \\ v_Y(x,y) \end{pmatrix} = \begin{pmatrix} x'-x \\ y'-y \end{pmatrix} \tag{1}$$

The affine motion model can be represented with the six parameters as follows:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_3 \\ a_6 \end{pmatrix} \tag{2}$$

In order to estimate the six affine motion parameters, we define an error function to be minimized by

$$E(a) = \sum_{i=1}^{N} \{[v_X(x_i,y_i) - \hat{v}_X(x_i,y_i)]^2 + [v_Y(x_i,y_i) - \hat{v}_Y(x_i,y_i)]^2\} \tag{3}$$

where N is the number of motion vectors in the same frame.

By substituting Eq. (2) into Eq. (3), we have

$$E(a) = \sum_{i=1}^{N} \{[V_X(x_i,y_i) - (a_1x + a_2y + a_3)]^2 + [V_Y(x_i,y_i) - (a1x + a2y + a3)]^2\} \tag{4}$$

The optimal values of the six parameters can be estimated by the least square method. The resulting equation is represented by

$$\sum_{i=1}^{N}\begin{bmatrix} x_i^2 & x_iy_i & x_i & 0 & 0 & 0 \\ x_iy_i & y_i^2 & y_i & 0 & 0 & 0 \\ x_i & y_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_i^2 & x_iy_i & x_i \\ 0 & 0 & 0 & x_iy_i & y_i^2 & y_i \\ 0 & 0 & 0 & x_i & y_i & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} v_X(x_i,y_i)\cdot x \\ v_X(x_i,y_i)\cdot y \\ v_X(x_i,y_i) \\ v_Y(x_i,y_i)\cdot x \\ v_Y(x_i,y_i)\cdot y \\ v_Y(x_i,y_i) \end{bmatrix} \quad (5)$$

C. Mosaic Background Generation

Once the affine parameters (M_1, M_2,...) have been calculated, we can warp all the images with respect to the common coordinate system. To create the final mosaic image, we map the transformation parameters for each frame into the reference coordinate system by concatenating the transformation matrices. We have arbitrarily selected the first image as the reference, and warped all other images into the first image's coordinate system. Using the camera motion information, a dynamic mosaic of the background is progressively integrated and stored in a single image. Pixel blending is used to reduce the discontinuities in color and in luminance.

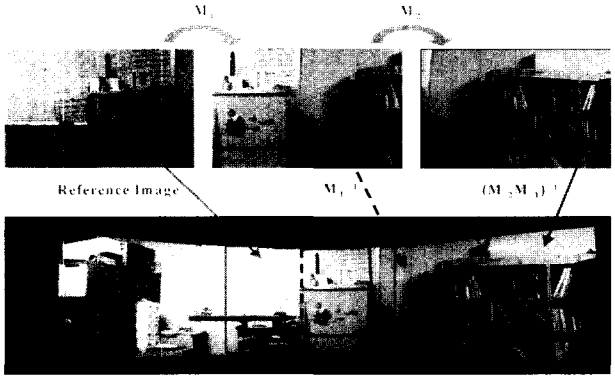


Fig. 3. Mosaic of a room from a sequence.

Fig. 3 shows an example of a sub-view in the mosaic image. The panoramic mosaic image is constructed from 6 views taken from equally spaced pan angle positions.

D. Object Detection

The mosaic image is to provide an initial rough reference background model for the background subtraction method. After the mosaic image is constructed, the live video captured from the camera is fed into the object detection module. The detection module monitors scene changes and trace the detected object when an intruding object is detected. The foreground object is extracted from the background and the extracted foreground is utilized as the basis to control the active camera to track the moving object. In addition, the separated background is utilized to update the corresponding background model to improve the segmentation result.

E. Object Tracking

Once the object region is detected, we can track the object efficiently by predicting the next coordinate from the observed coordinate of the object centroid. We design a 2D token-based tracking scheme using Kalman filtering [8].

The center position and the size of the moving object are used as the token $t(k)$ at time k. We assume the next token $t(k+1)$ is the sum of the current token $t(k)$ and the token change $\Delta t(k)$,i.e., we define a simplified polynomial motion model by

$$t(k+1) = t(k) + \Delta t(k) \quad (6)$$

In order to estimate the token change $\Delta t(k)$, we apply Kalman filtering on the system state $x(k)$, which is defined as a four-dimensional vector of the positional change of the target object per unit time interval and the size change of the target object.

$$x(k) = \begin{pmatrix} \Delta x\_center(k) \\ \Delta y\_center(k) \\ \Delta xsize(k) \\ \Delta ysize(k) \end{pmatrix} \quad (7)$$

The Kalman filtering algorithm estimates the system states based on a set of measurement errors. We assume that the state model is linear and defined by

$$x(k+1) = \Phi(k,k+1)x(k) + w(k) \quad (8)$$

$$\Phi(k,k+1) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (9)$$

where $x(k)$ denotes the system state at time k, $\Phi(k,k+1)$ denotes the state transition matrix in the unit time interval, and $w(k)$ denotes an estimation error vector. Assuming that the positional change and the size change of the target object per unit time interval are uniform, we can write the state transition matrix by Eq. (9).

We also assume a linear relationship between the system state and a set of measurements.

$$z(k) = H(k)x(k) + v(k) \quad (10)$$

$$H(k) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (11)$$

where $z(k)$ denotes a set of measurements, $H(k)$ an observation matrix, and $v(k)$ measurement errors.

Once we define the system model and the measurement model, we can apply a recursive Kalman filtering operation to obtain optimal linear minimum variance (LMV) estimates of motion parameters. As shown in Fig. 4, the recursive Kalman filtering algorithm consists of three main operations: initialization, state prediction, and measurement update.

The initialization step determines an initial state estimate $\hat{x}(0)$, an initial error covariance matrix $P(0)$ that represents the deviation of $\hat{x}(0)$ from the actual initial state $x(0)$, an autocorrelation matrix of estimation errors $Q(k)=E[w(k)w(k)^T]$, and an autocorrelation matrix of measurement errors $R(k)=E[v(k)v(k)^T]$. Initial values $\hat{x}(0)$ are derived by discrete-time derivatives of the object center locations and the sizes of the target object in the first two picture frames. We take the identity matrices for $P(0)$, $Q(k)$ and $R(k)$. After the initialization step, we switch to the tracking mode.

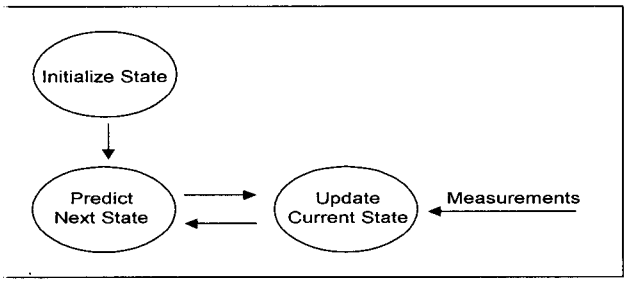


Fig. 4 Tracking operation.

The state prediction step determines a priori LMV estimate and its error covariance matrix for the current state based on the previous state estimate and its error covariance.

$$\hat{x}_{k+1}^{-} = \Phi_k \hat{x}_k \tag{12}$$

$$P_{k+1}^{-} = \Phi_k P_k \Phi_k^T \tag{13}$$

where $\hat{x}_{k+1}^{-}$ denotes a priori estimate for the system state at time $k+1$ based on measurements of $z_0, z_1, \ldots, z_k$ and $\hat{x}_k$ denotes an optimal estimate for the system state at time k based on measurements of $z_0, z_1, \ldots, z_k$. The measurement update step combines the estimated information with new measurements to provide an LMV estimate and its error covariance matrix for the current state.

$$\hat{x}_{k+1} = \Phi_k \hat{x}_k \tag{14}$$

$$P_{k+1}^{-} = \Phi_k P_k \Phi_k^T \tag{15}$$

$$P_{k+1}^{-} = \Phi_k P_k \Phi_k^T \tag{16}$$

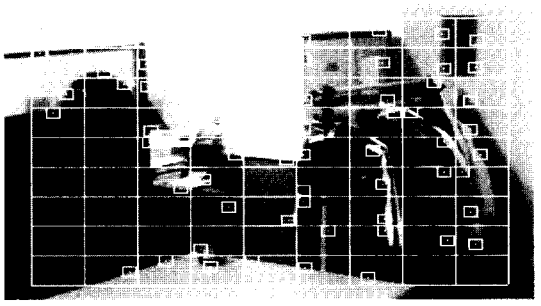
The term $K_k(z_k - H_k\hat{x}_k^{-})$ provides an optimal LMV estimate for $\hat{x}_k - \hat{x}_k^{-}$ based on z_k. It represents an optimal correction of the error incurred from the predicted estimate $\hat{x}_k^{-}$ of x_k. We perform this correction process with measurements on the positional change and the size of the target object.

The center position and the size of the object are used as the system states to be estimated. After we define the system model and the measurement model, we apply the recursive Kalman filtering algorithm to obtain linear minimum variance (LMV) estimates of motion parameters.

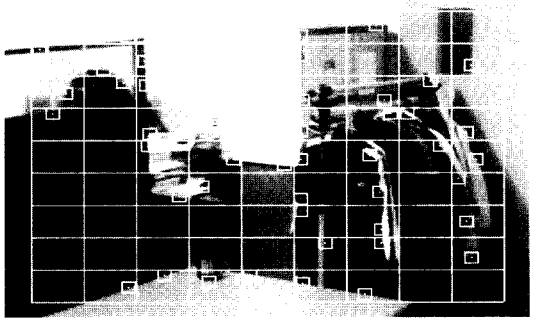
III. SIMULATION RESULTS

The proposed tracking system has been tested on several video sequences in indoor environments. Fig. 5 shows the block feature selection results for three activity thresholds. A high activity threshold diminishes the number of the block features. We use 35.0 as the threshold for the tracking system.

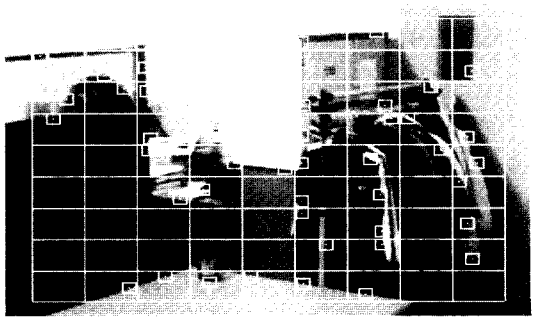
For types of sequences are captured as shown in Fig. 6; right-panning and left-moving person, right-panning and right-moving person, left-panning and right-moving person, left-panning and left-moving person. In the case of Fig. 6(a), the right panning of camera causes one motion. A moving person occurs the other motion. The background motion is separated by using dominant motion vector extraction. The center image of Fig. 6(a) displays the results of block motion vector estimation. The result of background motion separation is represented in the right image of Fig. 6(a).



(a)



(b)



(c)

Fig. 5 Block feature Selection for 3 Activity Thresholds: (a) TH(25.0) (b) TH(35.0) (c) TH(45.0).

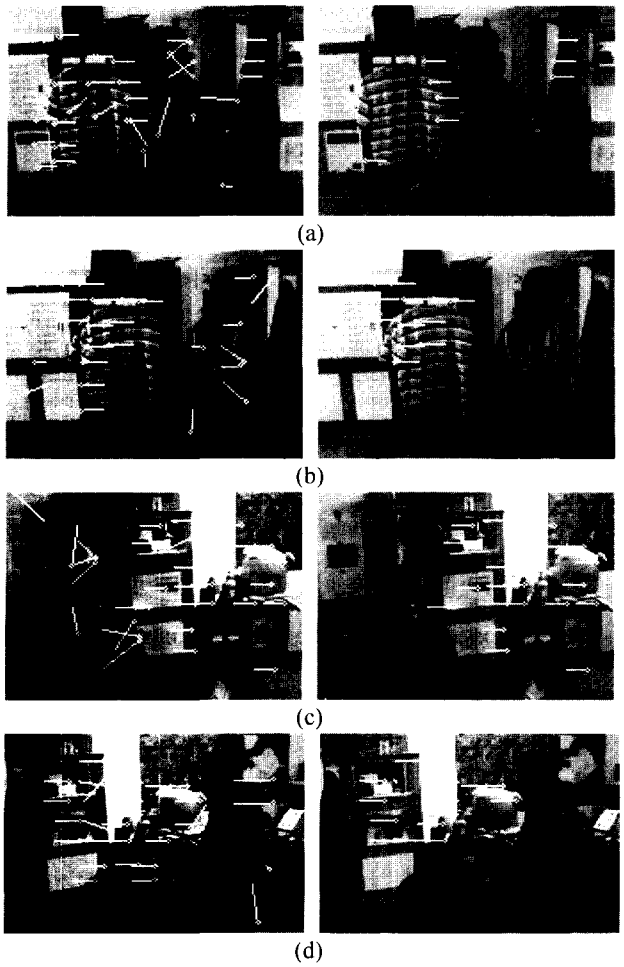


Fig. 6 Background Motion Separation: (a)right-panning and left-moving person, (b)right-panning and right-moving, (c)left-panning and right-moving person, (d)left-panning and left-moving person

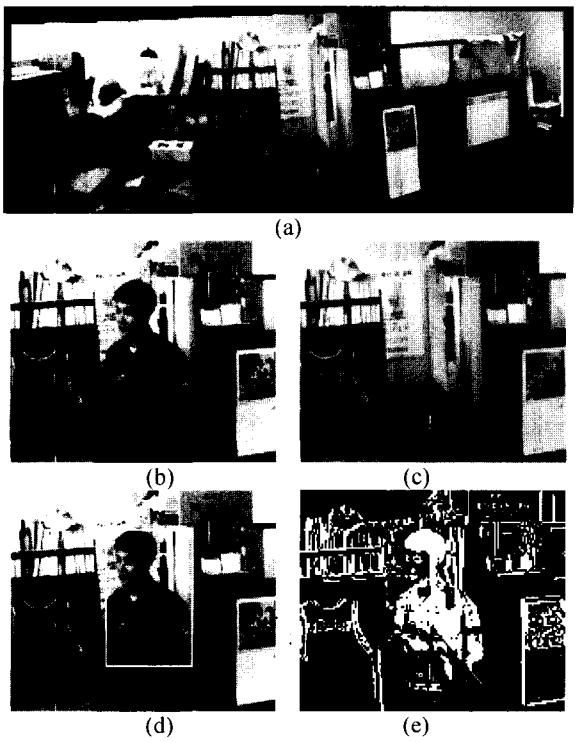


Fig. 7 Tracking Result I: (a) mosaic of a room from a sequence (b) current image (c) corresponding background image (d) extracted object (e) subtraction result (f) morphological opening (g) moving person extraction.

Fig. 7(a) shows the mosaic of a room from a sequence of 30 images. Fig. 7(b) and Fig. 7(c) show a current image with one person and the corresponding background from the mosaic. Fig. 7(e) is the subtraction image between the current image and the corresponding background image. This subtraction result has some noise blobs due to small errors motion We utilize a morphological opening operation to remove the noise blobs as shown in Fig. 7(f) and the largest blob is chosen to moving person in Fig. 7(g) and Fig. 7(d).

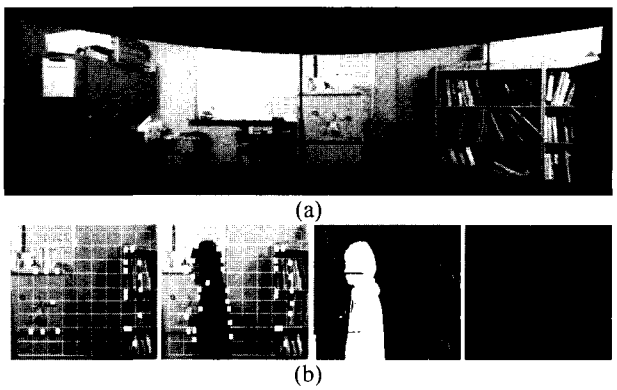


Fig. 8 Tracking Result II: (a) mosaic of a room from a sequence (b) moving person extraction.

Fig. 8(a) shows another tracking result. Fig. 8(b) shows a current image with one person and the corresponding background from the mosaic and shows the moving person extraction image between the current image and the corresponding background image.

IV. CONCLUSIONS

In this paper, we propose an algorithm for moving object tracking using image mosaic technique. An efficient camera motion estimation algorithm based on background motion is proposed to obtain a image mosaic integration. To build the mosaic, the frames are aligned with respect to a coordinate system and updated. Subtracting between current frame and the corresponding background region then segments the moving objects. As seen from the simulation results, the proposed algorithm successfully builds a background mosaic and segments the foreground objects.

REFERENCES

1] G. Johansson, "Visual Perception of Biological Motion and a Model for Its Analysis," Perception and Psycho-physics, Vol. 14, pp. 201-211, 1973.

2] K. Gould and M. Shah, "The Trajectory Primal Sketch", "A Multi-Scale Scheme for Representing Motion Characteristics," IEEE Conf. of CVPR, pp. 79-85, 1989.

3] O. Rouke and Badler, "Model-based Image Analysis of Human Motion using Constraint Propagation," IEEE Trans. on PAMI, Vol. 3, No. 4, pp. 522-537, 1980.

4] K.W. Lee, Y.H. Kim, J. Jeon and K.T. Park "An Algorithm of Moving Object Extraction Under Visual Tracking without Camera Calibration," Proceedings of ICEIC, pp. 151-154, 1995.

5] A.J. Lipton, H. Fujiyoshi and R.S. Patil, "Moving target classification and tracking from real time video," IEEE Workshop on Applications of Computer Vision, pp. 8-14, 1998.

6] R. Szeliski and H. Shum, "Creating full view panoramic image mosaics and environment maps," Computer Graphics Proceedings, pp. 251-258, 1997.

7] M. Irani, P. Anandan, J. Bergen, R. Kumar and S. Hsu, "Mosaic Representation of video sequences and their applications Signal Processing," Image Commnu., special issue on Image and Video Semantics: Processing, Analysis, and Application, Vol. 8, no. 4, pp. 673-676, 1996.

8] Y.K. Jung and Y.S. Ho, "Robust Vehicle Detection and Tracking for Traffic Surveillance," Picture Coding Symposium, pp. 227-230, 1999.

Young-Kee Jung

Received the B.S. degree in electric engineering from Seoul National University, Seoul, Korea, in 1986, and the M.S. degree in electric & electronic engineering from Korea Institute Science and Technology (KAIST), Taejon, Korea, in 1994, and the Ph.D. degree in Information and Communications Department of communications engineering from Kwangju Institute Science and Technology (K-JIST), Kwangju, Korea.

From 1986 to 1999, he was with LG Industrial System Laboratory, Anyang, Korea, where he was involved in development of machine vision and traffic surveillance system. Since 1999, he has been with Honam University, where he is currently Professor of Computer Engineering Department. His research interests include 3D image processing, object tracking and visual surveillance.

Dong-Min Woo

Received the B.S. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1980, and the M.S. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1982, and the Ph.D. degree in electric engineering from Case Western Reserve University, Cleveland, U.S.A.

From 1987 to 1990, he was with LG Industrial System Laboratory, Anyang, Korea, where he was involved in development of machine vision system. Since 1990, he has been with Meongji University, where he is currently Professor of Information Engineering Department. His research interests include image processing, stereo vision and 3D reconstruction.