JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 25, 1819-1837 (2009)

A Spatial-Extended Background Model for Moving Blobs Extraction in Indoor Environments*

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

This paper presents a system for extracting regions of moving objects from an image sequence. To segment the foreground regions of ego-motion objects, we create a background model and update it using recent background variations. Since background images are usually changed in blobs, spatial relations are used to represent background appearances, which may be affected drastically by illumination changes and background object motion. To model the spatial relations, the joint colors of each pixel-pair are modeled as a mixture of Gaussian (MoG) distributions. Since modeling the colors of all pixel-pairs is expensive, the colors of pixel-pairs in a short distance are modeled. The pixel-pairs with higher *mutual information* are selected to represent the spatial relations in the background model. Experimental results show that the proposed method can efficiently detect the moving object regions when the background scene changes or the object moves around a region. By comparing with Gaussian background model and the MoG-based model, the proposed method can extract object regions more completely.

Keywords: background modeling, background subtraction, object segmentation, video surveillance, mixture of Gaussian

1. INTRODUCTION

Human extraction is important in indoor surveillance systems for visual-based patient-caring, security-guarding, *etc*. In an indoor environment, people are usually considered to be the only foreground objects, which are defined as ego-motion objects. If the images with only background objects can be captured in advance, the positions of foreground objects can be detected by comparing the current image with the background images. However, background images vary when camera positions, background object positions, and illuminations change. Tracking objects in general environments will become very complicated.

In many surveillance applications especially in indoor environments, camera positions are generally fixed. Illumination variations and background object motion may

Received December 12, 2007; revised April 21, 2008; accepted May 8, 2008.
Communicated by H. Y. Mark Liao.
* This work was supported in part by the Ministry of Economic Affairs (MOEA), Taiwan, R.O.C., under grant No. 96-EC-17-A-02-S1-032, and in part by the Taiwan Information Security Center (TWISC), National Science Council under grant No. 95-2218-E-320-005.

change the captured images significantly. Examples of the motions include placing a book on a desk and moving a chair to another position. The positions of the objects are usually changed by people or other external forces. After the motion stops, these objects remain in the same position for a certain period; these motions are usually not repeated. In an indoor environment, the illumination of objects is not affected by continuous light changes, such as sun rise, sun set, or weather changes. Ignoring these continuous changes, brightness variations such as turning lights on or off, as shown in Fig. 1, and opening a window are assumed to be abrupt. Several researchers [2] assumed that brightness variations due to illumination changes are uniform. The right-hand side image in Fig. 1 shows the intensity differences after the lights were turned on. We observe that brightness variations in different pixels are not uniform. It is difficult to process this kind of variations. Several other researchers [2, 4, 5, 8] assumed that illumination changes are not repeated like the motion of freely movable objects. However, light sources can be repeatedly turned on or off several times over a period of time. The appearance changes on the illuminated regions will also be repeated.

Fig. 1. Left column: two consecutive images in different illumination conditions. Right image: intensity differences between the left two images.

To model the non-repeated background changes, we can use an *online updating* scheme to adapt to the background appearances in recently captured images [1-17]. When the appearances of a pixel repeatedly change, they can be modeled as a Mixture of Guassians (MoG) [9]. The online updating MoG model is useful for modeling rapidly repeated background appearances such as waves on water surface, but does not works well in long-term repeated appearances such as door opening and closing. In consecutive images, the repeated appearances of background objects usually appear in blobs in fixed places, while the appearances of foreground objects usually change their places and do not form fixed blobs. In this study, we extend the MoG model by using the *spatial relations* among pixels to model the background appearances.

The objective of our system is to extract moving object from a sequence of images. The system is divided into two modules: background modeling and foreground detection. The first module creates a background model to represent possible background appearances. The parameters of the model are learned and updated automatically from recently captured images. In the background model, the distributions of background features are

assumed to be mixtures of Gaussians [9]. Since background appearances are changed in blobs, the features used in the MoG should be able to represent spatial relations in the blobs. To represent the spatial relations, we estimate the joint color distributions of pixel-pairs in a short distance. Since estimating the distributions of all pixel-pairs is costly and not all pixel-pairs provide enough information to model background, we first calculate the dependence of colors in each pixel-pair. A pixel-pair with a higher color dependence implies that the two pixels provide more information to represent the appearance changes in blobs. Highly dependent pixel-pairs are then selected to model the spatial relation of background. In the second module, the background model that has already been updated from recent images is used to calculate the background probability of each pixel of the current image. The probability is then used to decide whether the pixel belongs to the foreground or background. Connected foreground pixels are extracted to form foreground regions.

The remainder of this paper is organized as follows. In section 2, we review related work on training methods of background models. In section 3, we describe the spatial-extended background model using the colors of pixel-pairs. In section 4, we design an algorithm to find the pixel-pairs that can provide more information to model spatial dependency. In section 5, we test the effectiveness of the background model in video clips and analyze the experimental results. In section 6, we conclude this paper.

2. RELATED WORK AND ISSUES IN BACKGROUND MODELING

A background model in a surveillance system represents background objects. The method that compares the current processed image with the background representation to determine foreground regions is called *background subtraction*. If the background is unchanged but affected by Gaussian noise, the colors of the background pixels can be modeled as a Gaussian distribution with mean vector (μ) and covariance matrix ($\sum$) [1-8]. Background subtraction is then performed by calculating the probability of each pixel in the current image belonging to the Gaussian model.

Since background appearances may be affected by external forces, modeling a pixel with a Gaussian distribution may misclassify some background pixels as foreground ones. In many cases, the background may change repeatedly. A background pixel with repeated changes can be divided into several *background constituents* and modeled as an MoG distribution [9-13]. For each background constituent in a pixel, the means (μ_i), covariances ($\sum_i$), and weights (w_i) of the ith constituent (b_i) have to be estimated. If there are K background constituents, the parameters of the background model can be represented as $\{\mu_i, \sum_i, w_i \mid 1 \le i \le K\}$. In order to decide whether a sample point X belongs to the background B, the conditional probability $P(B \mid X)$ is calculated as follows:

$$P(B \mid X) = \sum_{i=1}^{K} w_i P(b_i \mid X) \propto \sum_{i=1}^{K} w_i \eta(X; \mu_i, \Sigma_i), \tag{1}$$

where η represents a Gaussian probability density function,

$$\eta(X; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)}. \tag{2}$$

The motion of some background objects may not be repeated. After the motion, the objects remain in the same position for a period. To model the background changes, researchers have proposed methods for online updating of the parameters of background models [1-17]. The mean vector and covariance matrix in time t are represented as μ_t and $\sum_t$, respectively. The updating rules are formulated as follows:

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t, \tag{3}$$

$$\sum_t = (1 - \rho)\sum_{t-1} + \rho(X_t - \mu_t)(X_t - \mu_t)^T, \tag{4}$$

where ρ is used to control the updating rate. To integrate the updating method into an MoG model, Stauffer and Grimson [9] proposed a method to update the mean vector and covariance matrix of a background constituent to match those of X_t in Eqs. (3) and (4). The weight $w_{i,t}$ of the ith background constituent is updated as follows:

$$w_{i,t} = (1 - \alpha)w_{i,t-1} + \alpha I_{i,t}, \tag{5}$$

where $I_{i,t}$ is an indicator function, whose value is one if the ith background constituent matches X_t and zero otherwise, and α is a constant used to control the updating rate of the weights. In Stauffer and Grimson's method [9], the updating rate ρ for the parameters of the ith constituent (Gaussian distribution) is calculated according to α and $\eta(X; \mu_i, \sum_i)$.

To make background models more robust, researchers tried to modify updating rules or adopt different features [10-13]. In adaptive background models, background objects are assumed to appear more frequently than foreground ones. However, the assumption is not always satisfied. If the appearances of a background pixel appear less frequently than those of foreground objects, the background pixel is probably misclassified as a foreground object. Taking the following case as an example, assume a room is monitored by a fixed camera and the background objects in the room include a door and a wall as shown in Fig. 2 (a). People may enter the room, and close or open the door. If a person

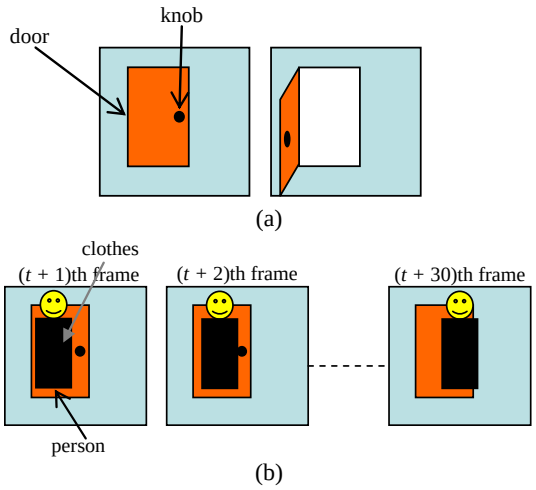
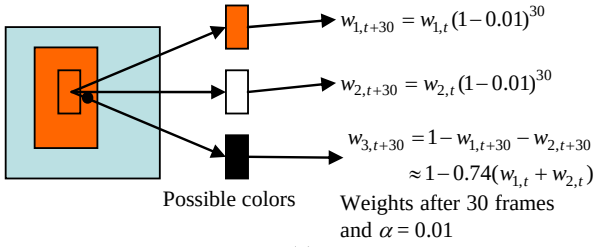
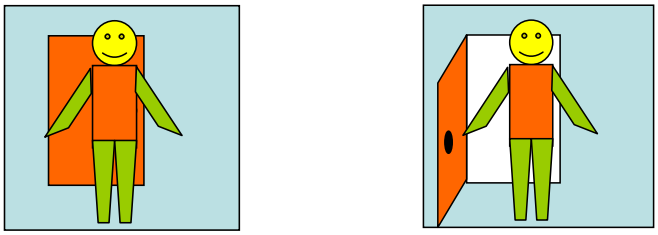


Fig. 2. (a) Sketches of two possible background scenes. Left: door closed; Right: door opened. (b) Consecutive frames of a person moving from left to right.



(c)

Fig. 2. (c) Possible colors and their weights of the rectangle region shown in the left image.



(a) Scene when the door is closed. (b) Scene when the door is opened.

Fig. 3. Sketches of a person whose clothes colors are similar to the door color in front of different background scenes.

wears a suit of clothes of single color and walks slowly across the room as shown in Fig. 2 (b), the major color of the clothes may be captured repeatedly in a certain position among several consecutive images. Assume that the person moves from left to right in 30 frames. If the updating rate α in Eq. (5) is set as 0.01, the weight of the repeatedly captured color of the clothes in the images will increase from 0 to 0.26, as shown in Fig. 2 (c). This large weight may cause the clothes to be labeled as the background, when the MoG model in Eq. (1) is used. Using a small updating rate can overcome this problem; however, the background model will be updated very slowly and may fail to learn background changes.

In another situation, the color of the person's clothes is assumed to be the same as that of the door, as illustrated in Fig. 3 (a). If the person enters the room and passes through the door, the region of clothes may be labeled as background due to the similarity of colors. However, after the door is opened, the current background color is not similar to the clothes color as shown in Fig. 3 (b). The clothes may still be labeled as the background, since they are very similar to possible background colors been estimated.

In these two situations, we observe that modeling each pixel independently cannot sufficiently represent the similarity among different object appearances caused by either object motion or illumination changes. Most researchers regarded the variations caused by object motion as foreground changes and attempted to eliminate the effect caused by illumination. In consecutive images, when the illumination changes, the pixels of an object are usually changed simultaneously. In order to model background objects, the pixels at different positions should be considered together. To represent the relation among the pixels, Durucan and Ebrahimi [14] proposed to model the colors of a region as vectors.

They segmented the foreground regions by calculating the linear dependence between the vectors of the current image and those of the background model. However, it is expensive to represent the dependence between vectors in terms of storage and speed. To reduce the cost, the vector of a region should be reduced to a lower dimensional feature. Li *et al*. [15, 16] used two-dimensional gradient vectors as the features of local spatial relations among neighboring pixels. In their proposed method, the appearance variations caused by illumination changes can be distinguished from object motion. However, the gradient features cannot be used to extract the foreground region that has a uniform color. To model the relation among pixels, we need to use the relations among near pixels to reduce time and storage consumption, and then extend the relations into a more global form.

The methods based on the Markov random field (MRF) are well known for extending the neighboring relations among pixels into a more global form. Image segmentation methods based on MRF [17, 19] assume that most pixels belonging to the same object have the same label and these pixels form a group in an image. The MRF combines colors among a clique of pixels in a neighboring system and uses an energy function to measure the color consistency. Then, the *maximum a posterior* estimation method is used to minimize the energy for all the cliques to find the optimal labels. In the MRF-based methods, the final segmentation results are strongly dependent on the energy functions of the labels in different cliques. If a high energy is assigned to the clique with unique labels, the extracted foreground regions will become more complete than those of the pixel-wise background models. An additional noise removal process is not required. However, when several pixels are mis-labeled, these errors will propagate into neighboring pixels. The error propagation will cause more pixels to be mis-labeled. In this research, we directly estimate the relations among pixels instead of the labels, and therefore the errors will not easily propagate.

3. JOINT BACKGROUND MODEL

In a sequence of images, colors will change in blobs instead of individual pixels due to illumination changes or object motion. This paper proposes to utilize the relations among pixels to represent the changes in blobs. The relations are formulated as a spatial-extended background model, which is then used to classify the pixels into either foreground or background.

3.1 Spatial Relation in Images

Using pixel-wise features, if the color of a foreground pixel is similar to those of the background, the pixel may be misclassified as background. If we can estimate the distributions of color combinations for the pixels in blobs, the foreground objects can be classified more precisely. Suppose there is a red door in a room and the appearance outside the door is white. When a person wearing a suit of interlaced red and white stripes passes through the door, parts of the suit may be misclassified as background when the colors of the pixels are modeled independently. Nevertheless, if we model the background appearances among pixels using joint multi-variate color distributions, the interlaced red and

white stripes can be classified as foreground using the method introduced later. However, estimating the multi-variate distributions for all pixel-pairs is still costly since the number of pixel combinations may be very large. In this research, we will estimate the color distributions of joint random vectors in closed pixel-pairs.

As stated in section 1, illumination changes and background object motions may change background appearances. Since the changes are complex, it is difficult to collect enough training samples for all the possible changes. In this paper, we modify Eq. (4) for updating the color distributions of pixel-pairs to adapt to the appearances that have not been trained, to be described in section 3.3.

3.2 Calculation of Background Probabilities

Assume that we have already estimated the color distributions of all background pixel-pairs. In this research, we decide whether pixel a_1 belongs to foreground according to its color and the color combinations of pixel-pairs (a_1, A), where A denotes a set of pixels associated with a_1.

Suppose that a sequence of pixels (a_0, a_1, …, a_n) has a corresponding color sequence (c_0, c_1, …, c_n). The probability of pixel a_0 belonging to background can be represented as $P(B_0 \mid x_0 = c_0, x_1 = c_1, \ldots, x_n = c_n)$, where the sequence ($x_0$, x_1, …, x_n) denotes the joint random variable of the colors for the sequence (a_0, a_1, …, a_n), and B_0 represents the event that pixel a_0 belongs to the background. Assuming that x_1, x_2, …, x_n are conditionally independent, based on the *naive Bayes' rule*, the probability $P(B_0 \mid x_0 = c_0, x_1 = c_1, \ldots, x_n = c_n)$ can be computed as the product of n pair-wise probabilities:

$$P(B_0 \mid x_0 = c_0, x_1 = c_1, ..., x_n = c_n) \approx P(B_0 \mid x_0 = c_0) \prod_{i=1}^{n} P(x_0 = c_0, x_i = c_i). \quad (6)$$

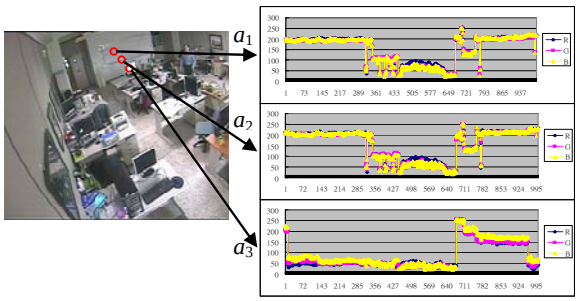
When estimating the background probabilities from above equation, we face two problems. The first one is the estimation and updating of the probability distributions $P(B_0 \mid x_0)$, $P(x_0, x_i)$, and $P(x_i)$. The distribution $P(B_0 \mid x_0)$, a pixel-wise background color distribution, is regarded as an MoG and can be calculated from Eq. (1), whose parameters are estimated and updated by using Eqs. (3)-(5). It is tedious to estimate and update the bivariate probability distribution $P(x_0, x_i)$, since the number of possible color combinations in x_0 and x_i is large. We will simplify the estimation and updating by combining the MoGs of pixels to form the joint random vector distributions of pixel-pairs. The second problem is the cost of modeling pixel-pairs. To model all pixel-pairs, the number of pixel-pairs is $O(W^2 \times H^2)$, where W and H are the width and height of the images, respectively. We reduce the complexity by only modeling the pixel-pairs that can provide sufficient information to represent spatial relations as described in section 4.

3.3 Estimation of Bivariate Color Distributions

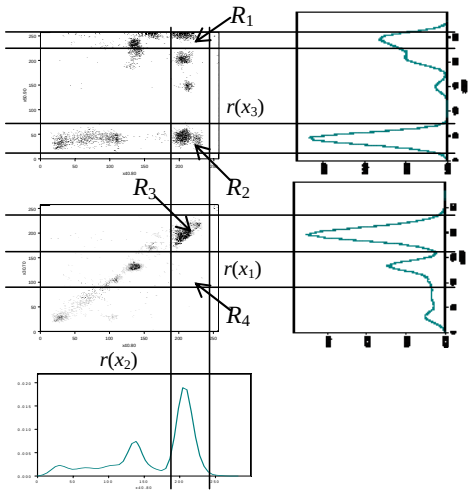
As mentioned before, the color distributions of pixel-pairs should be updated to adapt to the background changes. If we assume the color distributions in a pixel-pair (a_1, a_2) to be independent, the joint probability $P(x_1, x_2)$ can be regarded as $P(x_1) \cdot P(x_2)$. Assuming the color distributions are a mixture of Gaussians, the background colors of the

two pixels a_1 and a_2 form several *background constituents*, which can be represented as Gaussian distributions $G_1 = \{\eta(\mu_{k_1}, \sum_{k_1}) \mid 1 \le k_1 \le K_1\}$ and $G_2 = \{\eta(\mu_{k_2}, \sum_{k_2}) \mid 1 \le k_2 \le K_2\}$, respectively. The weights in both distributions are denoted as $W_1 = \{w_{k_1} \mid 1 \le k_1 \le K_1\}$ and $W_2 = \{w_{k_2} \mid 1 \le k_2 \le K_2\}$. When the independence is satisfied, the joint color of the pixel pair (a_1, a_2) forms $K_1 \times K_2$ *background joint constituents*, and the joint color distributions of the constituents are combinations of G_1 and G_2, denoted as $G = \{\eta(\mu_{k_1,k_2}, \sum_{k_1,k_2}) \mid 1 \le k_1 \le K_1, 1 \le k_2 \le K_2, \mu_{k_1,k_2} = (\mu_{k_1}, \mu_{k_2}), \sum_{k_1,k_2}$ is the covariance matrix $\}$. The weights of the joint constituents are $W = \{w_{k_1} \cdot w_{k_2} \mid 1 \le k_1 \le K_1, 1 \le k_2 \le K_2\}$. Since the parameters of G_1 and G_2 can be estimated from Eqs. (3) and (4), the parameters (G, W) of the bi-variate MoG $P(x_1, x_2)$ can be calculated easily.

In our background model, since the dependence between the colors of two pixels is used to model the spatial relations, the colors cannot be assumed independent. To estimate the parameter of a bi-variate MoG $P(x_1, x_2)$, we first examine the example depicted in Fig. 4. This figure shows the colors of three pixels a_1, a_2, and a_3 collected from 1,000 consecutive images, where a_1 and a_2 belong to the same object but a_3 does not. The



(a) A sample image and the colors in three pixels in a time period.



(b) Scatter plots of the pixels in the spaces $(r(x_2), r(x_3))$ and $(r(x_2), r(x_1))$, and probability distributions of $r(x_1)$, $r(x_2)$, and $r(x_3)$.

Fig. 4. Color samples of three pixels in 1,000 frames.

right-hand side image of Fig. 4 (a) shows the histograms of the colors in a_1, a_2, and a_3 in a time period of the sample image in the left. From the histograms, we observe that the colors of a_1 and a_2 usually change simultaneously and their values are dependent. The two scatter plots of Fig. 4 (b) from top to bottom are the scatters of ($r(x_2)$, $r(x_3)$) and ($r(x_2)$, $r(x_1)$), where $r(x_i)$ denote the red values of the color random variable x_i. The projection profiles from the top to bottom are the probability distributions of $r(x_3)$, $r(x_1)$, and $r(x_2)$. Each probability distribution forms several clusters and each cluster is regarded as a background constituent. As shown in the scatter plots, several combinations of background constituents in a pixel-pair form joint background constituents. When the probability distributions of a pixel-pair (a_1, a_2) are regarded as bi-variate MoGs, the probability function $P(x_1, x_2)$ is formulated as follows:

$$P(x_1 = c_1, x_2 = c_2) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} w_{k_1,k_2} \cdot \eta(c_{1,2}, \mu_{k_1,k_2}, \sum_{k_1,k_2}). \tag{7}$$

In the equation, the color vector $c_{1,2}$ is the joint color vector of colors c_1 and c_2, and the mean μ_{k_1,k_2} is the vector $[\mu_{k_1}, \mu_{k_2}]$, where μ_{k_1} and μ_{k_2} can be estimated from the background updating in Eq. (3). The covariance matrix $\sum_{k_1,k_2}$ is estimated with respect to the mean μ_{k_1,k_2} as

$$\sum\nolimits_{k_1,k_2}^{t} = (1-\alpha_{k_1,k_2}(c_{1,2}^t)) \sum\nolimits_{k_1,k_2}^{(t-1)} + \alpha_{k_1,k_2}(c_{1,2}^t)(c_{1,2}^t - \mu_{k_1,k_2}^t)(c_{1,2}^t - \mu_{k_1,k_2}^t)^T, \tag{8}$$

where the $c_{1,2}^t$ and μ_{k_1,k_2}^t are the joint color vector and joint mean vector in the pixel-pair (x_1, x_2), respectively. In the equation, if a joint vector of colors is matched with a joint constituent, the covariance matrix of the joint constituent should be updated as follows:

$$\alpha_{k_1,k_2}(c_{1,2}^t) = \begin{cases} \alpha_c, & \text{if } dist(c_{1,2}^t, \mu_{k_1,k_2}^t, \sum_{k_1,k_2}^t) < Th \text{ and} \\ & (k_1, k_2) = \text{argmin}(dist(c_{1,2}^t, \mu_{k_1,k_2}^t, \sum_{k_1,k_2}^t)), \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where α_c is a constant to control the updating rate, and $dist(c_{1,2}^t, \mu_{k_1,k_2}^t, \sum_{k_1,k_2}^t)$ is a distance function between the joint color vector $c_{1,2}^t$ and joint mean vector μ_{k_1,k_2}^t. The process of determining the minimal distance $dist(c_{1,2}^t, \mu_{k_1,k_2}^t, \sum_{k_1,k_2}^t)$ for all (k_1, k_2) is termed a *matching process*. In our experiments, the Mahalanobis distance is selected as the distance function. If a joint color vector $c_{1,2}^t$ does not match with any Gaussian distribution, a new Gaussian distribution is created and its mean is set as $c_{1,2}^t$. The weight of the new bi-variate distribution is initialized to zero.

The weight of a joint constituent in a pixel-pair is measured as the frequency of colors in the pixel-pair in past frames matched with the joint Gaussian distribution of the constituent, similar to Eq. (5). The updating rule of the weights is defined as

$$w_{k_1,k_2}^t = (1-\beta_{k_1,k_2}(c_{1,2}^t)) w_{k_1,k_2}^{(t-1)} + \beta_{k_1,k_2}(c_{1,2}^t), \tag{10}$$

$$\beta_{k_1,k_2}(c_{1,2}^t) = \beta_c \eta(c_1, \mu_{k_1}, \sum_{k_1}) \eta(c_2, \mu_{k_2}, \sum_{k_2}), \tag{11}$$

where β_c is a constant used to control the updating speed. Thus far, all the parameters used for estimating the joint color probability in Eq. (7) are ready and the background probabilities of Eq. (6) can be estimated from a set of color joint probabilities in a set of pixel-pairs.

Note that, during the background model estimation, the weight w_{k_1,k_2} is not set as the product of w_{k_1} and w_{k_2}. In other words, the constituents in the two pixels are not independent, and their relations are represented by the weights of the joint constituents. The relations in our model can be used to improve the accuracy of foreground detection. For example, in Fig. 4(b), the weight of joint constituents in R_4 is approximately zero, since no pixels match with the constituents; that is, the two constituents belonging to pixels x_1 and x_2 in R_4 usually do not appear simultaneously. However, when the joint colors in pixel-pair (x_1, x_2) match one of the joint constituents in R_4, the pixel-pair (x_1, x_2) is classified as foreground.

4. SPATIAL-DEPENDENT PIXEL-PAIRS SELECTION

The joint colors in pixel-pairs are used to represent the spatial relations of a background model. In a scene, not all pixel-pairs contain sufficient spatial relations. Modeling the unrelated pixel-pairs is useless for foreground detection. To reduce the computation cost, we first find the pixel-pairs with higher dependence.

The colors of two pixels with high dependence will form compact clusters in the scatter plots as shown in Fig. 4 (b). The compactness of a bi-variate distribution is measured from *mutual information* [20]. The mutual information $I(x_i, x_j)$ for colors c_i and c_j is defined as

$$I(x_i, x_j) = \sum_{\substack{\text{all } c_i \text{ in } x_i \\ \text{all } c_j \text{ in } x_j}} P(c_i, c_j) \log\left(\frac{P(c_i, c_j)}{P(c_i)P(c_j)}\right). \tag{12}$$

Here, $P(c_i)$, $P(c_j)$, and $P(c_i, c_j)$ can be computed from Eqs. (1) and (7). To reduce the cost of calculating the probabilities for all possible colors, the probability $P(c_i, c_j)$ can be replaced by the weights estimated from Eq. (10). The mutual information $I(x_i, x_j)$ in Eq. (11) can thus be reformulated as follows:

$$I(x_i, x_j) \approx \sum_{m=1}^{K_i}\sum_{n=1}^{K_j}\left(w_{m,n}\log\left(\frac{w_{m,n}\sum_{m''=1}^{K_i}\sum_{n''=1}^{K_j} w_{m'',n''}}{\sum_{n'=1}^{K_j} w_{m,n'}\sum_{m'=1}^{K_i} w_{m',n}}\right)\right). \tag{13}$$

The pixel pair (x_i, x_j) with higher mutual information $I(x_i, x_j)$ is selected to model spatial relations of the background model.

5. EXPERIMENTAL RESULTS

The test video clips used in our experiments are captured in two different sites and

by three different cameras. Two cameras are set in the two ends of a corridor (Cam1 and Cam2), and the other one is set in our laboratory (Cam3). The camera Cam1 is a gray-scale CCD camera, Cam2 a color CCD camera, and Cam3 a USB-Webcam. The resolution of each video frame is 320×240 and the frame rate is 30 fps. The total time of captured video clips is about 97.4 minutes, which include 175,394 frames. The clips contain moving humans, moving background objects, and changing illuminations.

In our experiments, we will compare the foreground detection results of three background models: the Gaussian background model (GBM), MoG-based model (MBM) [9], and spatial-extended background model (SBM). In both MBM and SBM, we represent the background color distributions as a mixture of six Gaussians. In our SBM, we model the pixel-pairs with the distance of five pixels, and use the two spatial-dependent pixel-pairs to represent the spatial relations.

To detect foreground, an effective background model can label most foreground pixels and very few background pixels as the foreground. The number of detected candidate foreground pixels is generally affected by the *foreground segmentation threshold* used to classify the pixels into foreground or background. Comparing different methods with unsuitable foreground segmentation thresholds cannot reflect the real performances. Here we perform two different kinds of experiments that detect foreground pixels via controlling either detected pixel numbers or foreground segmentation thresholds.

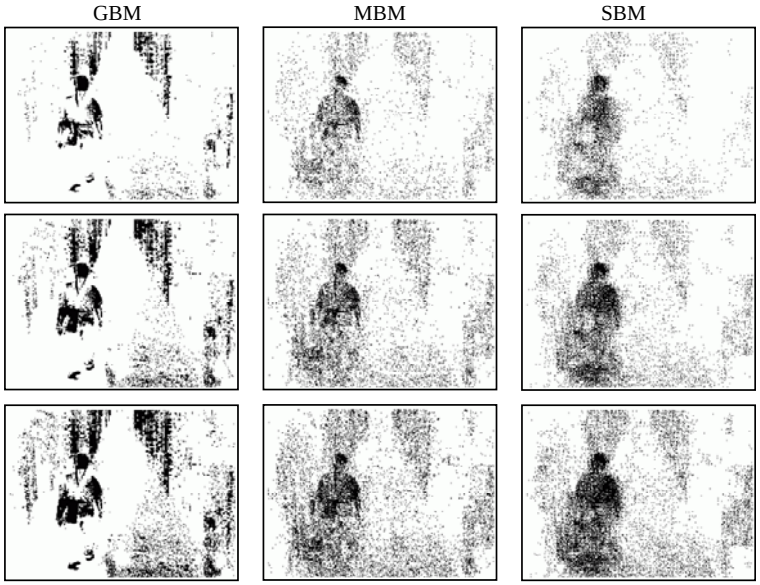
Figs. 5-7 show result images by controlling detected pixel numbers. Figs. 5-7 (a) show the original images, Figs. 5-7 (b) the detected foreground pixels using different methods, and Figs. 5-7 (c) the foreground regions of the images in the middle row of Figs. 5-7 (b) after morphology-based noise removal. In the noise removal process, if the closing operator is performed before opening near noises may be merged into a large one and cannot be removed by opening using the same structure element. If the opening is performed before closing, near small holes may not be removed. Therefore we first apply a closing operator with a smaller structure element (3×3) to fill the holes and then apply an opening with a bigger structure element (5×5) to remove noise pixels.

Fig. 5 (b) shows the results of a sample image captured by Cam1. Since the image is a gray scale one, different objects may easily have similar appearances. The distributions of joint random vectors of pixel-pairs are less efficient to distinguish different objects than those in color images. Thus, the results of SBM and MBM are much similar. After we apply noise removal as shown in Fig. 5 (c), the regions of the person using SBM are still more complete than those using MBM. The result shows our proposed SBM is better than the other two methods.

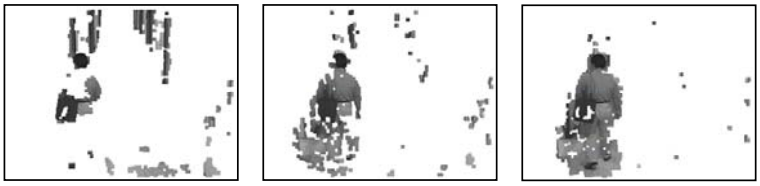
Fig. 6 (b) shows the results of a sample image captured by Cam2. The captured image is colored, and the colors of many parts of the person are similar to those of the background. The detected foreground regions of GBM and MBM are fragmental. Even though we apply a morphology-based hole filling procedure as shown in Fig. 6 (c), the foreground regions of the two methods are still fragmental. Thus, we can also conclude that SBM are more efficient than MBM and GBM.

Fig. 7 (b) shows foreground detection results of a sample image captured by Cam3. Some of the regions of the door and its shadow are misclassified as foreground ones by using GBM and MBM, but not misclassified by using SBM. In the sample, since the door is opened when the person enters the room, the GBM does adapt to the current appearance of the door and its shadow. In MBM, since the color distributions of the person,

(a) Original image.



(b) Detected foreground regions: the images from left to right are the results based on GBM, MBM, and SBM, and from top to bottom are results with 3*N*/40, 5*N*/40, and 7*N*/40 pixels. (*N* = image size).

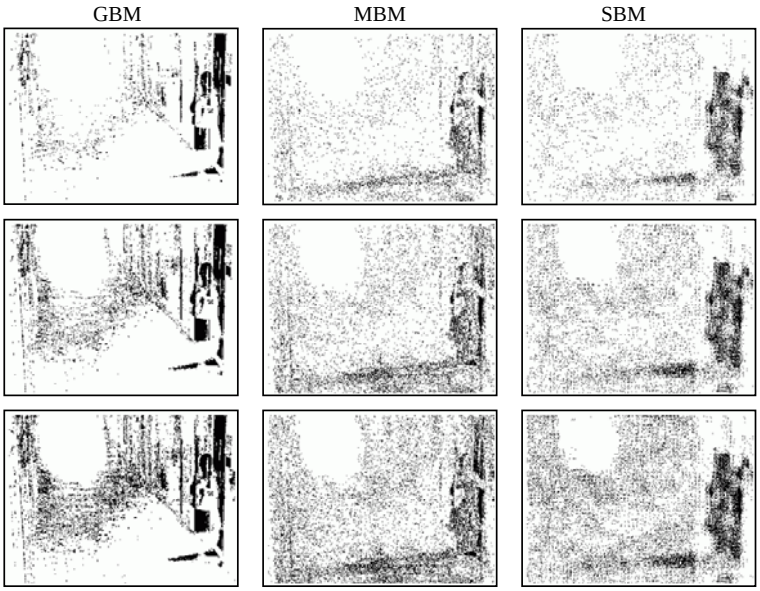


(c) Foreground regions after noise removal.

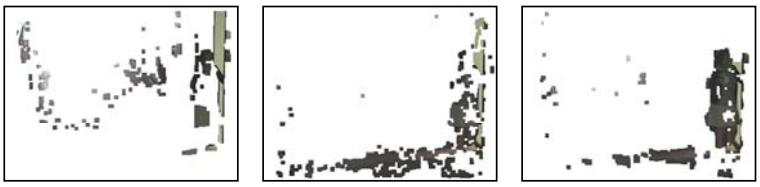
Fig. 5. Foreground detection results of an image captured by Cam1.

door and shadow may all be modeled, the regions of these object may be misclassified. As shown in the middle column of Fig. 7 (b), the regions of the door and its shadow may also be misclassified or those of the person may be fragmental when an unsuitable segmentation threshold is set. By adopting SBM, the appearances of the person are not taken as background, since the joint colors of a pixel-pair in the person are not captured repeatedly in the same position. The results in Fig. 7 (c) show that the person regions do not touch with background ones and less background regions are misclassified as foreground.

(a) Original image.

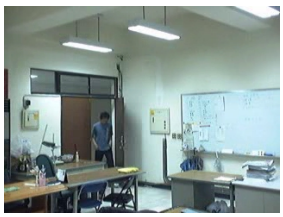


(b) Detected foreground regions: the images from left to right are the results based on GBM, MBM, and SBM, and from top to bottom are results with 3*N*/40, 5*N*/40, and 7*N*/40 pixels.



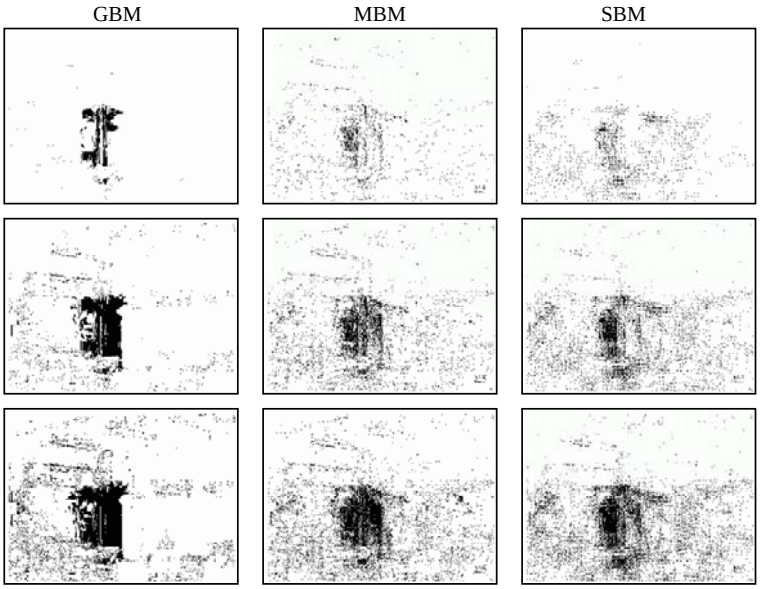
(c) Foreground regions after noise removal.

Fig. 6. Foreground detection results of an image captured by Cam2.

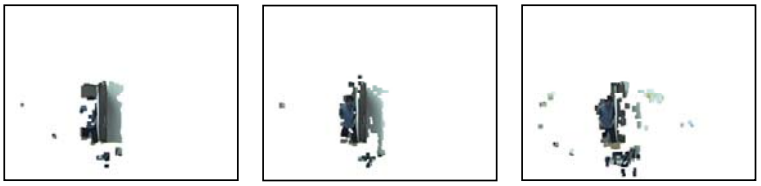


(a) Original image.

Fig. 7. Foreground detection results of an image captured by Cam3.



(b) Detected foreground regions: the images from left to right are the results based on GBM, MBM, and SBM, and from top to bottom are results with $N/40$, $3N/40$, and $5N/40$ pixels.



(c) Foreground regions after noise removal.

Fig. 7. (Cont'd) Foreground detection results of an image captured by Cam3.

Fig. 8 shows the foreground detection results of the images captured by the three cameras by setting a fixed threshold. The threshold that results in 15% false positive rate in training images is selected to test the performance of the models. The foreground regions depicted are noise removed. The results show that the foreground regions extracted by SBM are more complete, and the false positive regions are less than those of the other two methods. The persons in Figs. 8 (a), (c) and (f) walk around a place. Since the appearances of the persons are repeated in similar locations, the colors of the persons will be learnt as background by the pixel-wise background models GBM and MBM. In SBM, the colors of pixel-pairs will be modeled and the pixel-pairs without higher spatial dependency will be eliminated. Even though the appearances of a person are similar in a specific location, the joint colors of a pixel-pair in a fixed distance are usually varied and have low probabilities to be labeled erroneously as background. Note that the illuminations in these scenes are dramatically changed in Figs. 8 (e) and (f), when the lamplight is turned on, and slowly changed in Figs. 8 (a)-(d), when the illuminations are affected by the sunlight. In such environments, our proposed method is less affected by the illumination variations than others.

Fig. 8. Foreground detection results of the images captured by the three cameras. The images from top to bottom are original images, the results of GBM, MBM, and SBM.

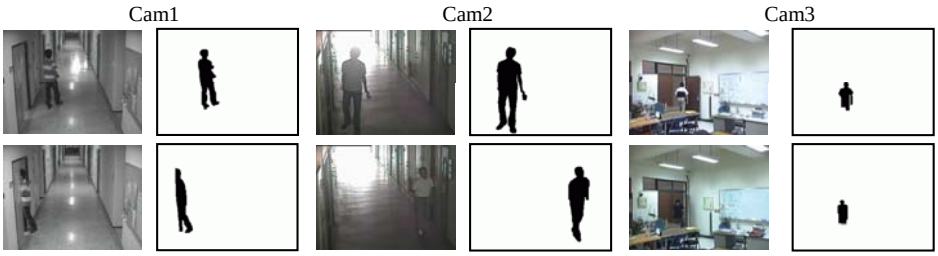


Fig. 9. Test samples and the manually labeled ground truth masks used for estimating the ROC curves.

Figs. 10-12 show the receiver operating characteristic (ROC) curves of the video clips by controlling thresholds. The results of each figure are estimated from 20 randomly selected test images. These images all include moving persons. The ground truth data of the test samples are manually labeled as shown in Fig. 9. The results show that the curves of SBM and MBM are very similar and the true positive rates of SBM are usually higher than that of MBM. When we fix the true positive rate on 80%, the false positive rates are about 21% and 30% for the test images captured by Cam2 (Fig. 11) using SBM and MBM, respectively. The results show that we can eliminate about 30% (9% in 30%) misclassified non-foreground pixels by extending MBM with spatial relations.

Note that the performances of GBM are usually better than those of MBM and SBM for the samples captured by Cam1 and Cam2 when the false positive rate is lower than 15%. The reason is that the background appearances do not change frequently in the corridor. In the environment with less frequently changed background, a Gaussian distribution can easily model the color distribution. However, when the background changes, the performance of GBM may become unacceptable for a fixed threshold as shown in Fig. 8.

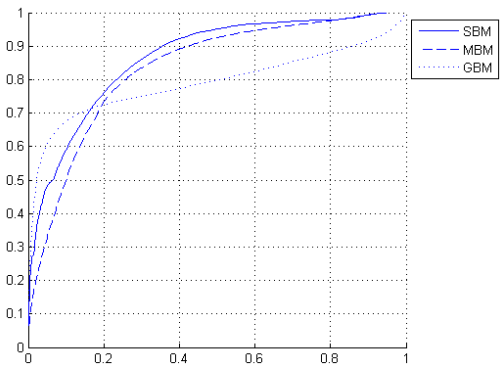


Fig. 10. The ROC curve of test images captured by Cam1.

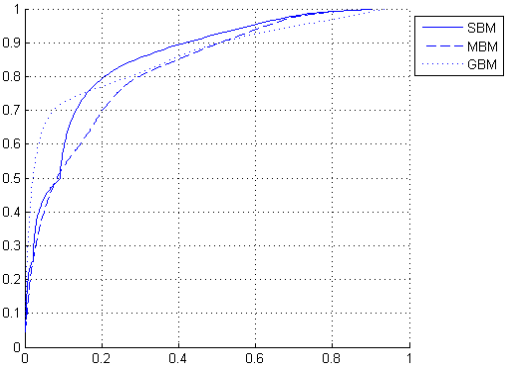


Fig. 11. The ROC curve of test images captured by Cam2.

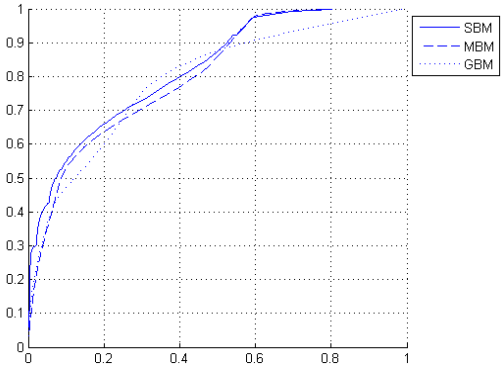


Fig. 12. The ROC curve of test images captured by Cam3.

Although experimental results show that SBM usually outperforms MBM and GBM, the SBM is slower than the other two methods. The computation complex of calculating the background probability of a pixel of MBM is $O(K)$, where K is the number of background constituents, but SBM is $O(K \times M)$, where M is the number of pixel-pairs used. When we update the model of a pixel, the computation complex of MBM is still $O(K)$, but SBM is $O(K^2 \times M)$ since the computation cost of updating each weighting matrix is $O(K^2)$. On a PC with Pentium4 2 GHz CPU, the SBM can perform about one frame per-second, but the MBM is about 10 frames per-second. In our tests, about 90% CPU time spends on calculating the mutual information (Eq. (13)) and updating the matrix w (Eq. (10)).

6. CONCLUSIONS

In this paper, we have proposed a system to extract moving object regions from consecutive images. Firstly, we have developed a spatial-extended background model for foreground detection. In the background model, we have used the probabilities of joint random vectors between near pixels to model the spatial relations. To reduce the cost of

modeling the pixel-pairs, we calculate the mutual information in each pixel-pair for finding the spatial-dependent pixel-pairs.

In general environments, when the background regions are stable, the Gaussian background model is suitable to segment foreground regions. However, when background regions change, the model is unsuitable. To detect foreground regions more accurately with respect to either changed or still background regions, we should combine our propose model with Gaussian background model. To achieve this, some heuristic rules should be created for deciding which model should be selected. This is left for future studies.

REFERENCES

1. S. Shiry, Y. Nakata, T. Takamori, and M. Hattori, "Human detection and localization at indoor environment by home robot," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 2, 2000, pp. 1360-1365.
2. L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, Vol. 36, 2003, pp. 585-601.
3. A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, Vol. 90, 2002, pp. 1151-1163.
4. 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.
5. C. 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.
6. S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, Vol. 80, 2000, pp. 42-56.
7. A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, 2003, pp. 918-923.
8. R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, 2003, pp. 1337-1342.
9. C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 745-757.
10. D. S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, 2005, pp. 827-832.
11. H. Wang and D. Suter, "A re-evaluation of mixture-of-Gaussian background modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, 2005, pp. 1017-1020.
12. K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, Vol. 11, 2005, pp. 172-185.

13. D. R. Magee, "Tracking multiple vehicles using foreground, background and motion models," *Image and Vision Computing*, Vol. 22, 2004, pp. 143-155.
14. E. Durucan and T. Ebrahimi, "Change detection and background extraction by linear algebra," *Proceeding of the IEEE*, Vol. 89, 2001, pp. 1368-1381.
15. L. Li, W. Huang, I. Y. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Transactions on Image Processing*, Vol. 13, 2004, pp. 1459-1472.
16. L. Li and M. K. H. Leung, "Integrating intensity and texture differences for robust change detection," *IEEE Transactions on Image Processing*, Vol. 11, 2002, pp. 105-112.
17. Y. Wang, T. Tan, K. F. Loe, and J. K. Wu, "A probabilistic approach for foreground and shadow segmentation in monocular image sequences," *Pattern Recognition*, Vol. 38, 2005, pp. 1937-1946.
18. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, New York, 2001.
19. H. Deng and D. A. Clausi, "Unsupervised image segmentation using a simple MRF model with a new implementations scheme," *Pattern Recognition*, Vol. 37, 2004, pp. 2323-2335.
20. C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, Vol. 14, 1968, pp. 462-467.
21. S. L. Zhao and H. J. Lee, "Human silhouette extraction based on HMM," in *Proceedings of the 18th International Conference on Pattern Recognition*, Vol. 2, 2006 pp. 994-997.

San-Lung Zhao (趙善隆) received the B.S. and M.S. degrees in Computer Science and Information Engineering form National Chiao Tung University, Hsinchu, Taiwan, in 1998, and 2000, respectively. He is currently a Ph.D. student in Computer Science and Information Engineering, National Chiao Tung University, Taiwan. 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). Since Aug. 1998, he had been the Chief Secretary to the president. 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, he has been the Dean of Academic Affairs. 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.