Estimation of Number of People in Crowded Scenes Using Perspective Transformation

Sheng-Fuu Lin, *Member, IEEE,* Jaw-Yeh Chen, and Hung-Xin Chao

***Abstract*—In the past, the estimation of crowd density has become an important topic in the field of automatic surveillance systems. In this paper, the developed system goes one step further to estimate the number of people in crowded scenes in a complex background by using a single image. Therefore, more valuable information than crowd density can be obtained. There are two major steps in this system: recognition of the head-like contour and estimation of crowd size. First, the Haar wavelet transform (HWT) is used to extract the featured area of the head-like contour, and then the support vector machine (SVM) is used to classify these featured area as the contour of a head or not. Next, the perspective transforming technique of computer vision is used to estimate crowd size more accurately. Finally, a model world is constructed to test this proposed system and the system is also applied for real-world images.**

***Index Terms*—Crowd density, crowd size, perspective transform.**

I. Introduction

RECENTLY, efforts in crowd estimation at exhibition centers, stadiums, airports, and subways have been addressed in the research field of automatic surveillance systems [1]–[5]. The estimation of crowd size is a difficult problem because in a crowd, only parts of people's bodies appear. As crowd density increases, the overlap among crowd members gets worse. Moreover, there are significant varieties in color and texture of the crowd, and the backgrounds against which the people lie are unconstrained and complex. An estimating system of crowd size should overcome all of the above challenging problems and work well and robustly.

For real-time estimation of crowd density, two systems in London [1] and Genova [4]–[6] have been proposed based on existing installed closed circuit television (CCTV). The two systems basically employ a number of sample image processing techniques for feature extraction over the image frames from CCTV. The image processing techniques are summarized as follows. First, background removal, an idea which is also used in [7] and [8], is used to measure the area occupied by the crowd versus that of the background. Edge detection is an alternative idea to measure the total perimeter of all the regions occupied by people.

For extracting significant features, the technique described in [9]–[12] estimates crowd densities uses the gray level dependence matrix (GLDM) method [13] to carry out texture analysis and a neural network (NN), implemented according to the Kohonen's self organizing map (SOM) model, for the task of crowd density estimation. Marana *et al.* [10] presented a technique for automatic crowd density estimation based on *Minkowski fractal dimension* of the image of the area under monitoring.

As for designing an algorithm to classify and estimate the number of people in crowds, there are many kinds of methods that could be used. In [14], the crowd classification is performed by the hybrid global learning (HGL) algorithm which combines the least-squares method together with different global optimization methods, such as random search (RS), simulated annealing (SA), and genetic algorithm (GA). For detecting a human head, the concept of NN-based face detection [15] might be considered. However, not all people will directly face the camera; therefore, the reverse side, right side, or left side views of people in a crowd also will need to be observed in an image. In this case, the contour of the human head would be a better choice than the human face for detecting people in a crowd. Papageorgiou *et al.* [16] used a Haar wavelet representation to capture the structural similarities between instances of an object class such as a pedestrian, and the support vector machine (SVM) [17] is used for classification in these papers.

Based on the literature, it can be seen that there are many different methods to estimate crowd density. These researchers inspire the motivation of this work. First, most research of these related papers is based on the estimation of crowd density. The estimated result of their research is usually the density of a crowd as proposed by Polus *et al.* [18]. However, judgments of crowd density are usually different for every person. Therefore, the alternative idea of estimating the number of people in crowds is proposed in this paper. Second, for certain purposes, simplifying the problem or increasing computation speed, the system is restricted in that it is aimed at only a fixed local area to estimate crowd density. When the estimating system is set up some place, it first has to capture a reference image which does not contain any people, and some parameters need to be modified. These systems lack generality and robustness. The model developed in this paper can work with any background without a reference image, as will be illustrated in the following sections.

II. Crown Size Estimation System

The structure of the system is shown as a block diagram in Fig. 1. There are two parts in the flowchart for this system. The left side of the flowchart is the *model constructing phase*, and the right side is the *testing phase*.

The objective in the model constructing phase is to construct the *support vector classifier* which can determine whether or not

Manuscript received September 29, 2000; revised March 26, 2001 and August 30, 2001. This paper was recommended by Associate Editor V. Murino.

The authors are with the Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan, R.O.C.

Publisher Item Identifier S 1083-4427(01)10800-3.

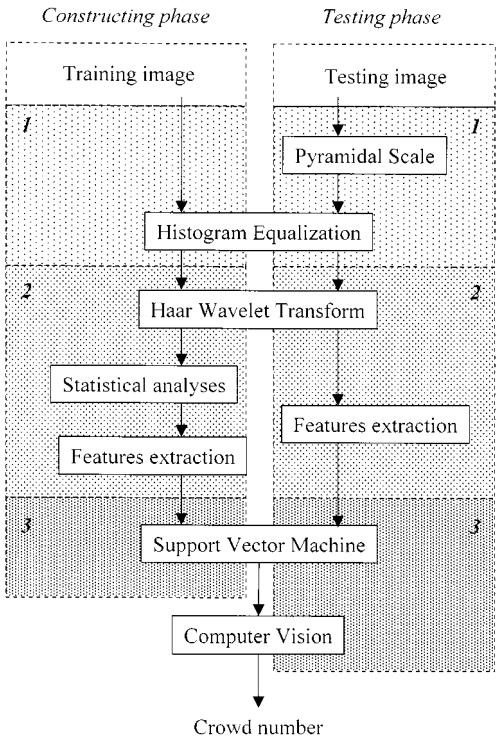


Fig. 1. Block diagram of the system.

the input features are the contours of people. This phase involves three stages:

1) image preprocessing;
2) features extracting;
3) support vector classification.

For minimizing the influence caused by different illumination, the histogram equalization is used first to preprocess the input image. Second, to get the significant features of all training images, the images are processed through Haar wavelet transform (HWT) and statistical analysis, as shown in Section II-A. Then these features are used to train the system.

The testing phase shows the complete procedure of how the crowd size is estimated from the beginning. To achieve multiscale (human head) detection, the image is iteratively resized before the histogram equalization and then processed by the next stage. After all possible locations of people are detected by the previously constructed model, the information of the sizes and positions of these detected frames are used to estimate the number of people in crowds with a perspective transformation method (also called imaging transformation).

A. Feature Extraction

In general circumstances, only the heads of people can be observed when there is a large crowd. Thus, the contour of the human head is chosen as the detected target. As the crowd increases, the size of a human head in a fixed picture decreases. Therefore, to achieve the goal of estimation of crowd size, the size of detected template of the human head is 16×16 (the smallest detectable size of the contours of people) in this paper. To develop the model for the class of head-like contour, a set of 1030 gray raw images (obtained from photos) of the human head of size 16×16 is used. Using the Haar wavelet representation, both the coarse-scale (2×2 pixels) and fine-scale (4×4 pixels) features are used. At these scales of wavelets, there are 1182 [i.e., $3\times (13 \times 13 + 15 \times 15)$] total features for a 16×16 pattern.

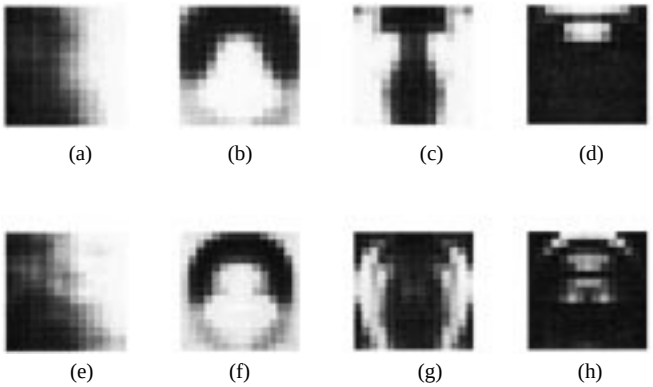


Fig. 2. Ensemble average values of the wavelet coefficients coded using gray level. (a) Average coefficients of random scenes with scale (4×4). (b) Average, (c) vertical, and (d) horizontal coefficients of images of people with scale (4×4). (e) Average coefficients of random scenes with scale (2×2). (f) Average, (g) vertical, and (h) horizontal coefficients of images of people with scale (2×2).

To extract the significant coefficients of these features, they must be analyzed using statistical method. The basic analysis in identifying the important coefficients consists of the following steps. First, the wavelet coefficients of each input pattern are normalized to minimize the influence of different input pattern. Second, the average of normalized coefficients along the ensemble is calculated Third, the standard deviation of all the coefficients based on its corresponding average is computed. Finally, these results are sorted and analyzed to select the important coefficients.

Consider the wavelet coefficient c_{ij}^k. The normalization step represented by the following equation:

$$\mathbf{C}_{ij}^k = \frac{c_{ij}^k}{\max\{c_{ij}^k\}}, \qquad 0 \le \mathbf{C}_{ij}^k \le 1 \tag{1}$$

$j = 1, 2, \ldots, P$ for any i and k, where P is the number of coefficients in an input image, c_{ij}^k is the jth coefficient of the ith input image in the kth coefficient class, and $\mathbf{C}_{ij}^k$ is the normalized wavelet coefficient which is bounded between 0 and 1. To normalize the coefficients is to preserve the relation in distribution of gray level that the scale of gray level will be ignored. The average jth coefficient for each class k, $\overline{\mathbf{C}_j^k}$ is calculated by

$$\overline{\mathbf{C}_j^k} = \frac{\sum_{i=1}^{Q} \mathbf{C}_{ij}^k}{Q} \tag{2}$$

where Q is the number of input images.

As shown in Fig. 2, a gray level coding scheme is used to visualize the average patterns in the different classes of coefficients. Note that the raw images are too numerous to show in Fig. 2. The gray level coding for the arrays of coarse scale coefficients (4×4) are shown in Fig. 2(a)–(d). The gray level coding for the arrays of finer scale coefficients (2×2) are shown in Fig. 2(e)–(h). Fig. 2(a) shows the vertical coefficients of random images which do not contain a human head, and this figure is uniformly gray as expected. The corresponding images, which

are not shown here, of the average, horizontal, and diagonal coefficients, are similar. On the other hand, the coefficients of the human head, shown in Fig. 2(b)–(d), show clear patterns with the different classes of wavelet coefficients being tuned to different types of structural information. The average wavelets, shown in Fig. 2(b), show the entire characteristics of a human head. The vertical wavelets, shown in Fig. 2(c), obtain both the right and left sides of the human head. The horizontal wavelets, shown in Fig. 2(d), capture the top and chin of the head. The wavelets of finer scale in Fig. 2(e)–(h) provide better spatial resolution of the human head, and overall shape and smaller scale detail appear clearer. To select the most significant features, the standard deviation of all of the coefficients based on its corresponding average is computed here. The standard deviation of human head coefficients is far below the standard deviation of these nonhead coefficients. When the standard deviation of a coefficient is low, it means it is more reliable and it has a higher relativity importance.

Finally, the data according to the importance of their relativity is sorted and there are 36 (out of 1182) significant coefficients to be selected. There are six average, three horizontal, and six vertical coefficients at the scale of 2×2, and eight average, seven horizontal, and six vertical coefficients at the scale of 4×4. These coefficients serve as the feature vector for the classification problem.

B. System Training

The classification technique chosen here is the SVM [17]. This recently developed technique possesses the special advantage of having very few tunable parameters and of using structural risk minimization which minimizes a bound on the generalization error. To train the system, a database of images of human heads captured at arbitrary visual angles with complex backgrounds and images which do not contain any people as the negative examples are used. Virtually any image which does not contain any people can serve as a nonhead example because the space of nonhead images is much larger than the space of human head images. However, collecting a representative set of nonhead is difficult as there are no typical examples of nonheads. To overcome the problem of defining this extremely large negative class, a bootstrapping training is used. As shown in Fig. 3, the initial positive and negative training sets are used to train this system. Then, those false detection of patterns will be added to the database of negative examples and the classifier is then retrained with this larger set of data. By using these iterations of the bootstrapping procedure on this system, the classifier can be constructed to more completely identify nonhead class.

C. Vanishing Point

In this paper, three restrictive conditions are assumed in the crowd size estimation model. It is assumed that all of the frames of human heads have the same size in the real world, all crowds distribute over a horizontal plane, and the center of image is equal to optical center. These three conditions help to simplify the estimating problem. All three-dimensional (3-D) lines with a nonzero slope along the optic axis have perspective projections

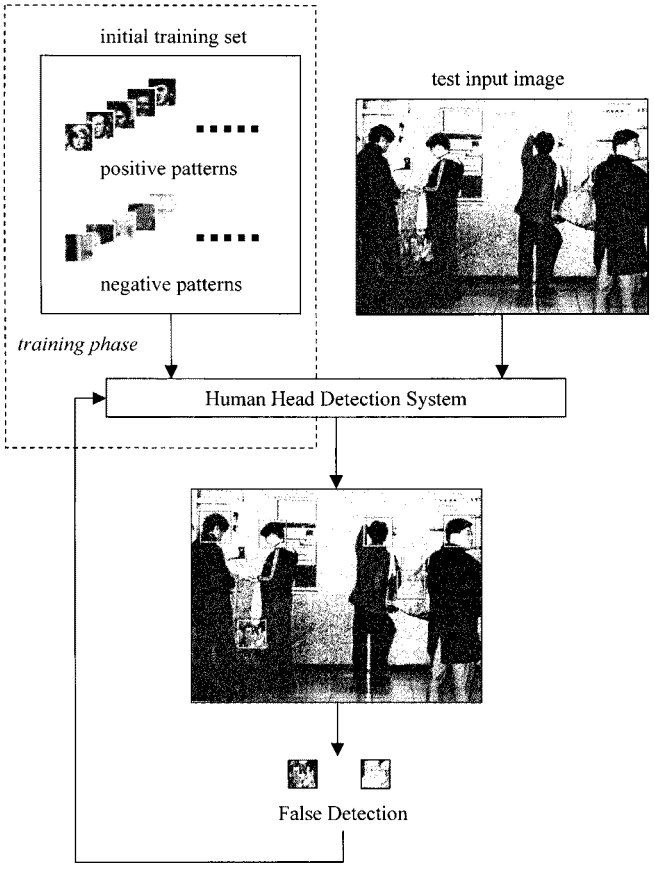


Fig. 3. Framework of bootstrapping training.

that meet at the same point, called the *vanishing point*, on the perspective projection image plane [19].

Let visual angle θ_c be the angle between the plane of camera sensor and the horizontal plane. Furthermore, let l be the focal length of the CCD camera and lv be the distance between the vanishing point and the center of the CCD sensor. The model of the vanishing point formation process is illustrated in Fig. 4. It can be easily observed that the position of the vanishing point is relative to θ_c. Therefore, θ_c can be calculated if the vanishing point is found out first.

After the frames of human heads are detected (to detect different frame sizes, the images will be resized before the histogram equalization), those frames of a different size are rearranged on the vertical center line of the image. To find out the position of the vanishing point, all top-left points of these frames are connected with a straight line. However, because of the error due to the precision of the classifier and the assumed conditions, these points could not be perfectly connected by one straight line. Thus, the approximate line is calculated by linear regression [20]. Before the method is used, those frames with different sizes at the same row are averaged to find a representative frame. Therefore, the approximate line could be more correct. After the optimal linear function is calculated, the vanishing point can be obtained on the point of intersection which is intersected by the regression line and the center line in the image. Fig. 5(a) shows an example where there are 39 frames of human heads which are detected by SVM, and Fig. 5(b) shows the result of the vanishing point as calculated by linear regression.

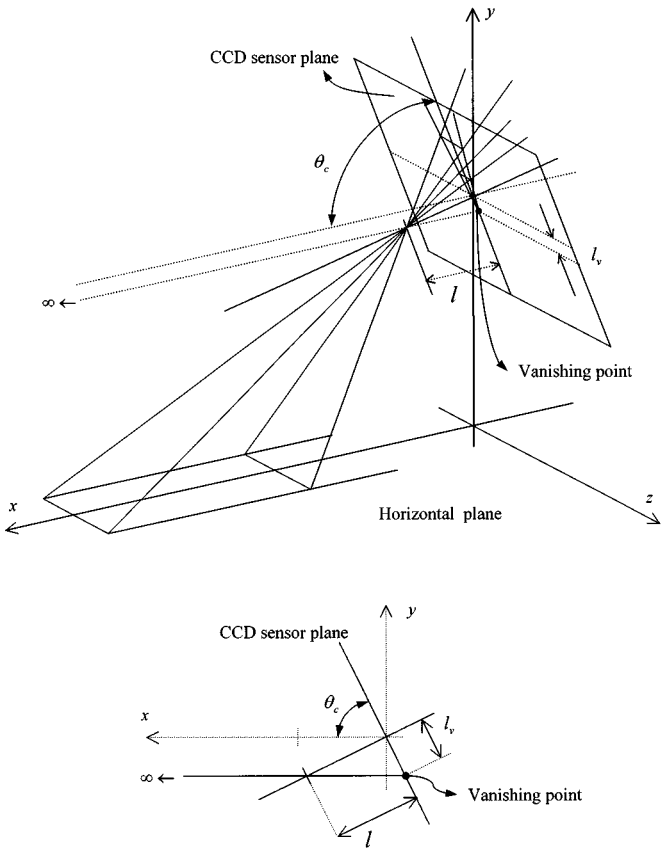
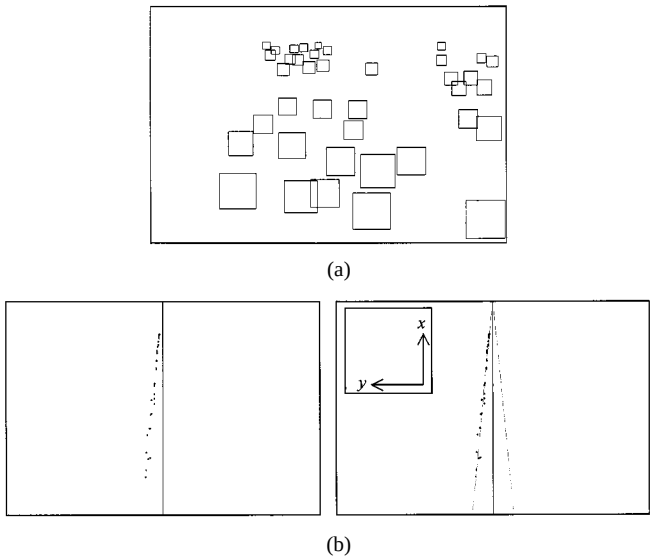


Fig. 4. Model of the vanishing point formation process.



(a)

(b)

Fig. 5. (a) Example where there are 39 frames of human heads which are detected by SVM. (b) Vanishing point which is calculated by linear regression. Linear function is $y = -0.095\,135x + 44.7735$ where the original point is at the bottom of the vertical center line.

D. Equidistant Parallel Lines in Computer Vision

After the position of the vanishing point is decided, the visual angle θ_c can be calculated by

$$\theta_c = \arctan \frac{l}{l_v} \tag{3}$$

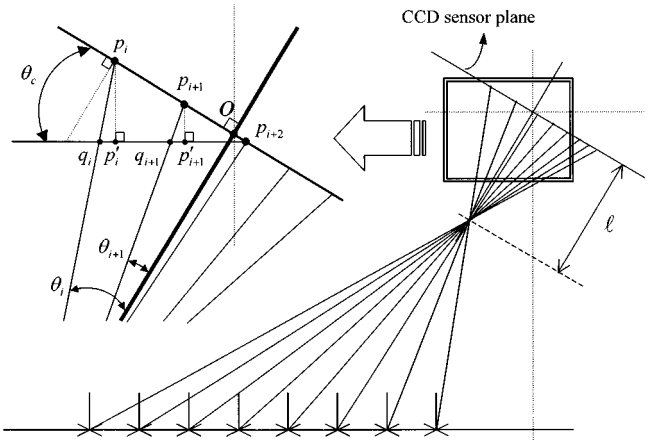


Fig. 6. Model of the projection of a number of parallel lines.

where $0 \le l_v \le \infty$, and $0° \le \theta_c \le 90°$. Suppose there are several equidistant parallel lines which are parallel to the CCD sensor on the horizontal plane in the real world. As θ_c changes, these equidistant parallel lines will show the different form of the density distribution in the image.

Referring to Fig. 6, the distribution of those equidistant parallel lines could be drawn by the following equation:

$$\theta_i = \arctan\left(\frac{\overline{op_i} \cdot l_p}{l}\right) \tag{4}$$

$$\theta_{i+1} = \arctan\left(\frac{\overline{op_{i+1}} \cdot l_p}{l}\right) \tag{5}$$

where l_p is the real length of a pixel in the image, and

$$\overline{p_{i+2}q_i} = (\overline{p_ip_{i+1}} + \overline{p_{i+1}p_{i+2}}) \cdot [\cos\theta_c + \sin\theta_c \cdot \tan(\theta_c - \theta_i)] \tag{6}$$

$$\overline{p_{i+2}q_{i+1}} = \overline{p_{i+1}p_{i+2}} \cdot [\cos\theta_c + \sin\theta_c \cdot \tan(\theta_c - \theta_{i+1})]. \tag{7}$$

By using similar triangles, the following relation can easily be found:

$$\overline{p_{i+2}q_i} : \overline{p_{i+2}q_{i+1}} = 2 : 1. \tag{8}$$

Substituting (6) and (7) into (8), the segment $\overline{p_{i+1}p_{i+2}}$ can be calculated. To provide more details, the possible situations are divided into three different cases. There are four variables which are defined for simplifying the expression of the equation first

$$\begin{aligned} I_0 &= \cos\theta_c + \sin\theta_c \cdot \tan(\theta_c - \theta_i) \\ I_1 &= \cos\theta_c + \sin\theta_c \cdot \tan(\theta_c + \theta_i) \\ I_2 &= \cos\theta_c + \sin\theta_c \cdot \tan(\theta_c - \theta_{i+1}) \\ I_3 &= \cos\theta_c + \sin\theta_c \cdot \tan(\theta_c + \theta_{i+1}). \end{aligned} \tag{9}$$

Then, the representations of $\overline{p_{i+1}p_{i+2}}$ can be divided into the following cases.

Case 1: If $\overrightarrow{op_i} < 0$ and $\overrightarrow{op_{i+1}} \le 0$, then the segment $\overline{p_{i+1}p_{i+2}}$ can be calculated by

$$\overline{p_{i+1}p_{i+2}} = \overline{p_ip_{i+1}} \cdot \frac{I_0}{2I_2 - I_0}. \tag{10}$$

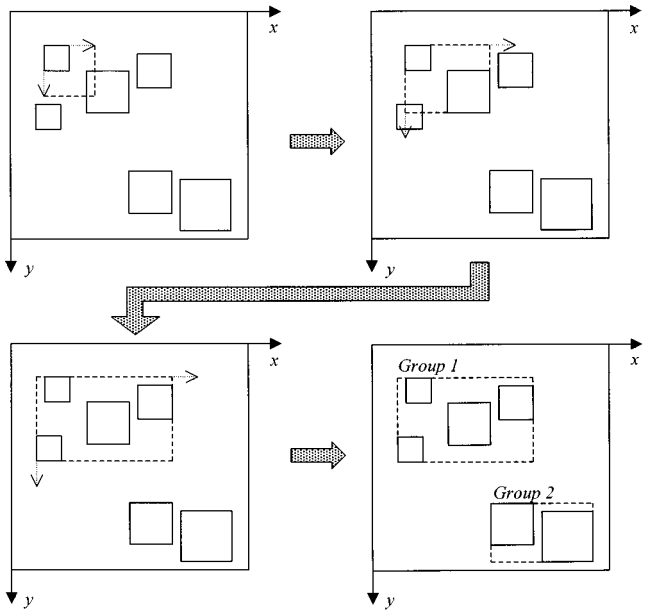


Fig. 7. Procedure of the grouping.

Case 2: If $\overrightarrow{op_i} < 0$ and $\overrightarrow{op_{i+1}} > 0$, then the segment $\overline{p_{i+1}p_{i+2}}$ can be calculated by

$$\overline{p_{i+1}p_{i+2}} = \overline{p_ip_{i+1}} \cdot \frac{I_0}{2I_3 - I_0}. \tag{11}$$

Case 3: If $\overrightarrow{op_i} \geq 0$ and $\overrightarrow{op_{i+1}} > 0$, then the segment $\overline{p_{i+1}p_{i+2}}$ can be calculated by

$$\overline{p_{i+1}p_{i+2}} = \overline{p_ip_{i+1}} \cdot \frac{I_1}{2I_3 - I_1}. \tag{12}$$

Finally, the distance of the projection of those parallel lines can be calculated. The information is significant for estimation of the number of people in crowded scenes.

E. Estimation of Crowd Size

In this estimating model, it is first assumed that the arrangement of crowds is general and the distance between people is equidistant. Here, an effective method is proposed to estimate the number of people in crowds and get an acceptable testing result. The estimation algorithm is described as follows.

Because the distribution of a crowd may not only be a concentric group, a simple method which can divide these frames into several groups is used here if the distributions of crowd are dispersed concentrically. In order to clarify the process, Fig. 7 illustrates a simple example. Each frame could be a group of crowd. For the top-left point, which is the starting point for scanning each frame, both the lines in x and y directions are extended to enclose a region known as the scanning region. If there is another frame intersected in the scanning region, this frame is regarded as a member belonging to the same group. Then, the scanning region is expanded to a new scanning region which contains these two frames and the top-left point of this region is treated as the starting point again. This process is repeated until the quantity of the extension in x and y directions are both equal to the threshold value which will be defined later, and the scanning region still does not contain any other frames.

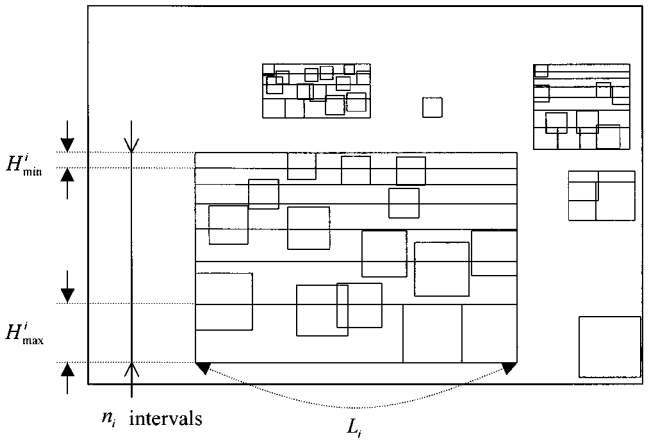


Fig. 8. Definition of some variables during estimating process.

Therefore, all of the frames of human heads can be divided into several groups of crowds. As the size of frames changes, the condition of formation of the crowded group should be different. Therefore, the threshold value of the quantity of the extension in x and y directions is variable. The threshold value T is defined as

$$T_i = \alpha \cdot L_i, \qquad i = 0, 1, 2, \ldots, n \tag{13}$$

where n is the number of all of the frames in the image, L_i is the length of the side of ith frame, and α is the grouping coefficient, selected here as 0.5.

Next, since both the gradient of the arrangement and the distribution of the crowd groups are known, the approximate number of people in crowds can be estimated by the relation between the grouping area and total area of frames with different sizes. As mentioned earlier, all groups of crowds are assumed so that all people are arranged in compliance with the same global rule. According to this rule, to estimate the density of crowd and the number of people will be the utmost that can be done. However, this rule means that the estimating result may not be correct because every person in each crowd group may not actually arrange with the global rule.

For obtaining a better estimating number, the above-mentioned assumption is modified. The local gradient parallel lines of each group region is computed separately to calculate the number of people in a crowd. As shown in Fig. 8, let n_i be the number of intervals which is divided by the projection of parallel lines in ith crowd group. Let L_i be the broad length of the scope of ith crowd group, $H^i_{\min}$ be the length of the minimum frame in ith crowd group, and $H^i_{\max}$ be the length of the maximum frame in ith crowd group. Then, the approximate number of people in ith crowd group could be estimated by the following:

$$P_i = \sqrt[n_i]{\frac{H^i_{\min}}{H^i_{\max}}} \tag{14}$$

where P_i is the decreasing parameter for reducing the size of the frame of the human head

$$\begin{aligned} A &= \sum_{i=0}^{m-1} A_i \\ &= \sum_{i=0}^{m-1} \sum_{j=0}^{n_i-1} a_{ij} \\ &= \sum_{i=0}^{m-1} \sum_{j=0}^{n_i-1} \frac{L_i}{H^i_{\max} \cdot P^j_i + H^i_{\max} \cdot S \cdot P^j_i} \\ &= \sum_{i=0}^{m-1} \sum_{j=0}^{n_i-1} \frac{L_i}{H^i_{\max} \cdot P^j_i \cdot (1+S)} \end{aligned} \tag{15}$$

where

- A total number of people in the image;
- A_i number of people in ith crowd group;
- a_{ij} number of people at jth interval in ith crowd group
- S diluted parameter that considers the space between human heads.

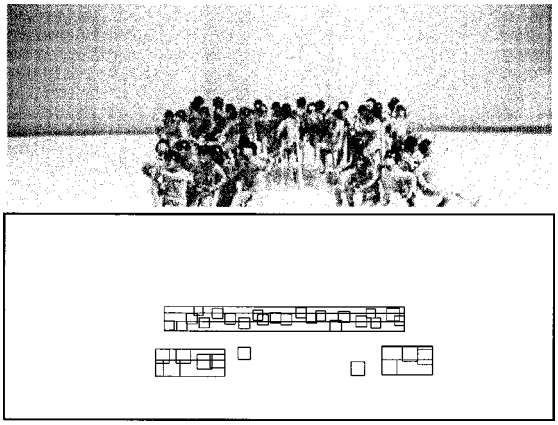
In general, if the crowd density tends to high, the diluted parameter is about $1.5 \sim 3$; if the crowd density tends to low, the diluted parameter is about $3 \sim 5.5$. Finally, the number of people in crowds is calculated. In the given example, the estimated crowd size is 58. For verifying the accuracy of the proposed algorithm, many conditions with different levels of complexity will be tested in the next section.

III. Experimental Results

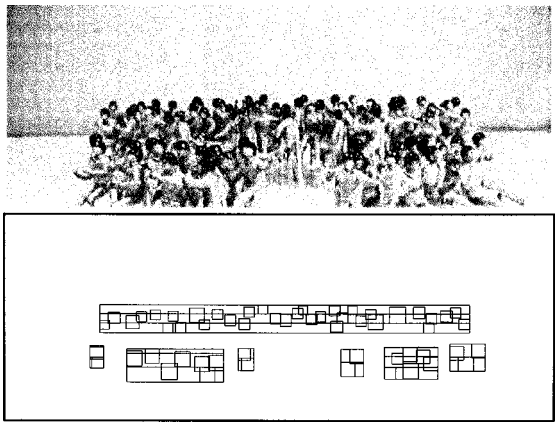
To verify the performance of the proposed system without controversy, the correct number of people in crowds should be provided in order to be compared with the result estimated by the system. A model world was constructed for this purpose. In this model world, there are 125 person-like puppets that are used to simulate a crowd in the real world. Then, many sizes of crowds and visual angles between the plane of the camera sensor and the horizontal plane are regulated to generate many test images. All of the information such as the number of people in crowds and visual angle is recorded correctly in advance to be compared with results estimated by the system. An evaluation of the performance of the proposed model is presented in the experiments of different numbers of crowd in the model world and the real world separately. After all of the parameters are determined off-line, four pictures can be dealt per second by the proposed algorithm (where a personal computer with Pentium III, 800 MHz processor is used). All of these experiments and an analysis will be detailed in the following section.

A. The Results of Estimation of Crowd Size in the Model World

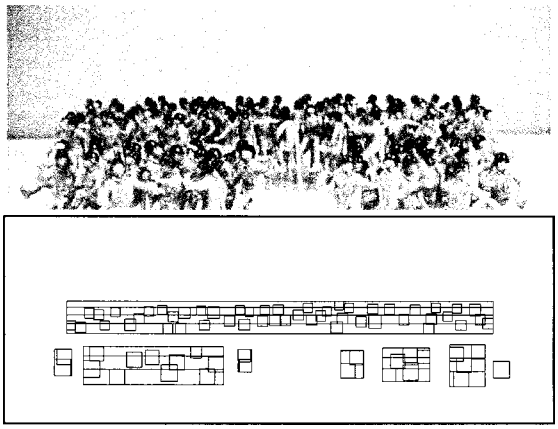
First, the images that contained different levels of crowd density with simple backgrounds in the model world were tested. In this experiment, the crowd is constructed by a number of puppets in a model world and the number of the puppets can be controlled and well known. Furthermore, the testing images can be captured at arbitrary visual angles to make the testing procedure more complete. Therefore, there is an absolute and correct number of puppets in the crowd that is used as the right answer to compare the result for every testing image. To probe the performance of the proposed system, several kinds of conditions were simulated and experimented upon. The angle between the plane of camera sensor and the horizontal plane θ_c, the size of the crowd group N_g, the crowd size N_a, and the crowd density are the four influential factors to this system which are discussed. The upper image in each set of images is the testing image, and the lower image is obtained as the result of processing the testing image. For each image, two numbers are shown in the caption: the exact number of puppets in the image and the final



(a)

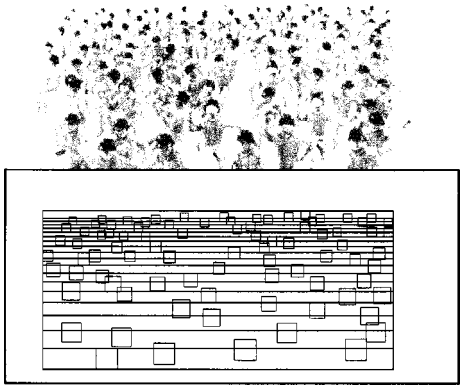


(b)

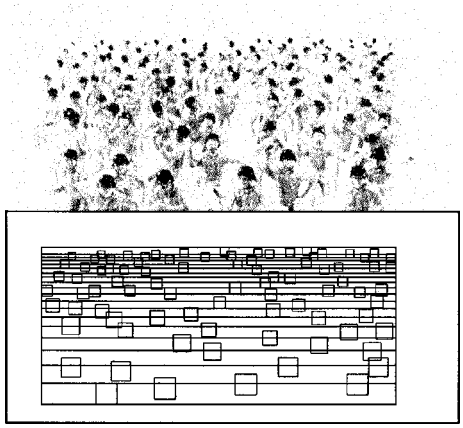


(c)

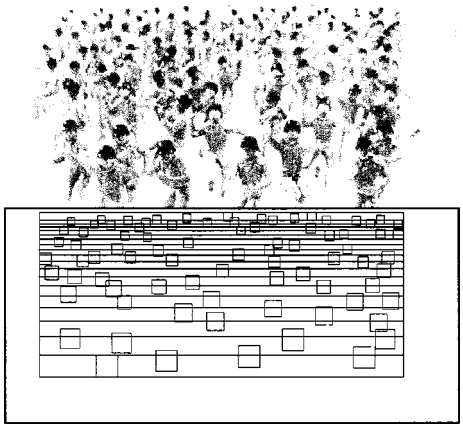
Fig. 9. Three kinds of estimated examples. Both the parameters $\theta_c = 72.5°$ and $N_g = 3$ are fixed, and each crowd size is separately $N_a = 40,\ 80,$ and 120 in series. (a) 40/40. (b) 80/78. (c) 120/117.



(a)



(b)



(c)

Fig. 10. Three kinds of estimated examples. The parameters $N_a = 110$ is fixed, $N_g =$ random and angles are separately $\theta_c = 72.5°$, 68.5°, and 64.5° in series. (a) 110/110. (b) 110/109. (c) 110/112.

estimated number of puppets. For example, a set of two numbers (110/100) means that there are 110 puppets in the image and 100 puppets finally estimated. Fig. 9 shows three kinds of crowd sizes ($N_a = 40$, 80, and 120) that have been divided into three crowd groups ($N_g = 3$) at a fixed angle ($\theta_c = 72.5°$) and tested. Fig. 10 shows the fixed crowd size ($N_a = 110$) with random crowd groups ($N_g =$ random) at three kinds of angles ($\theta_c = 72.5°$, $\theta_c = 68.5°$, and $\theta_c = 64.5°$). These tested results show that the crowd size can be estimated close to the real number even when the head detection is not very good.

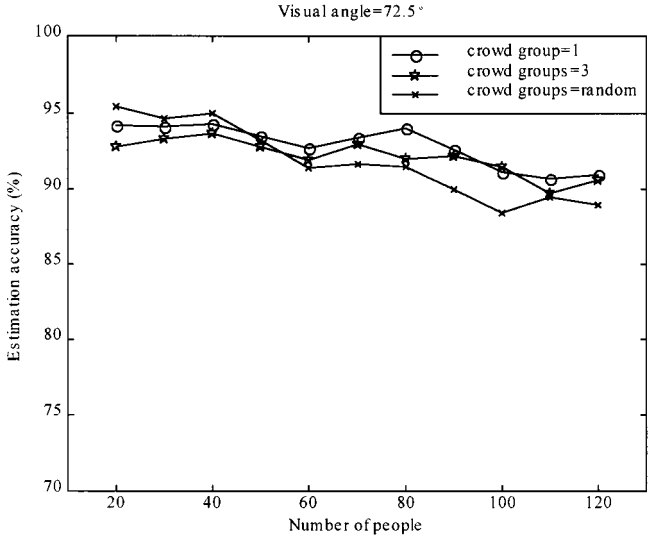


Fig. 11. Plot of the estimation accuracy of different crowd sizes and crowd groups with fixed visual angle.

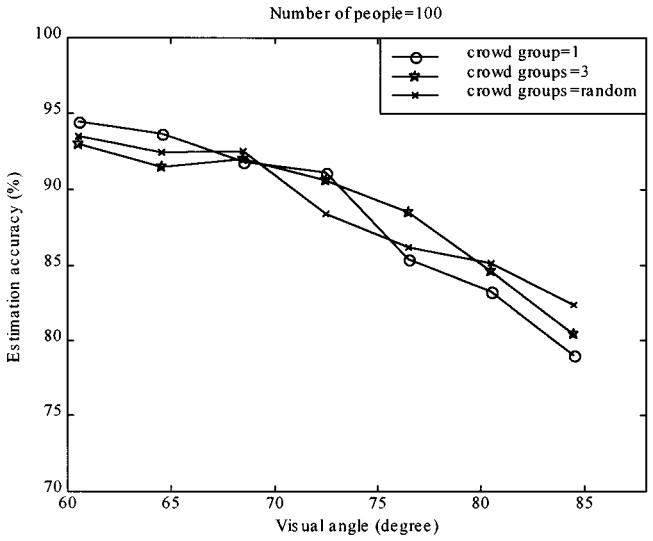


Fig. 12. Plot of the estimation accuracy of different visual angles and crowd groups with fixed crowd size.

In the following, the performance of the system under different conditions is illustrated in Figs. 11 and 12. In each case, we test 250 samples, calculate the absolute error and divide it by the right number to obtain the error rate. Then, accuracy is 100% minus the error rate. In Fig. 11, the relationship between the estimation accuracy, the crowd size, and the crowd group at the same visual angle can be observed. When the number of the crowd group is small and the distribution of people in crowd is concentrated, the estimation accuracy tends to be high; if not, it tends to be low. This is because when the size of crowd group is fewer and the distribution is more concentrated, the estimating conditions are nearer the assumed conditions of the system. Even though the accuracy decreases as the crowd size increases, the overall accuracy is still around 90%∼95%. In Fig. 12, the relationship between the estimation accuracy, the visual angle, and the crowd group at the same crowd size is described. It can be clearly observed that the estimation accuracy of the system is strongly affected by the visual angle θ_c. When

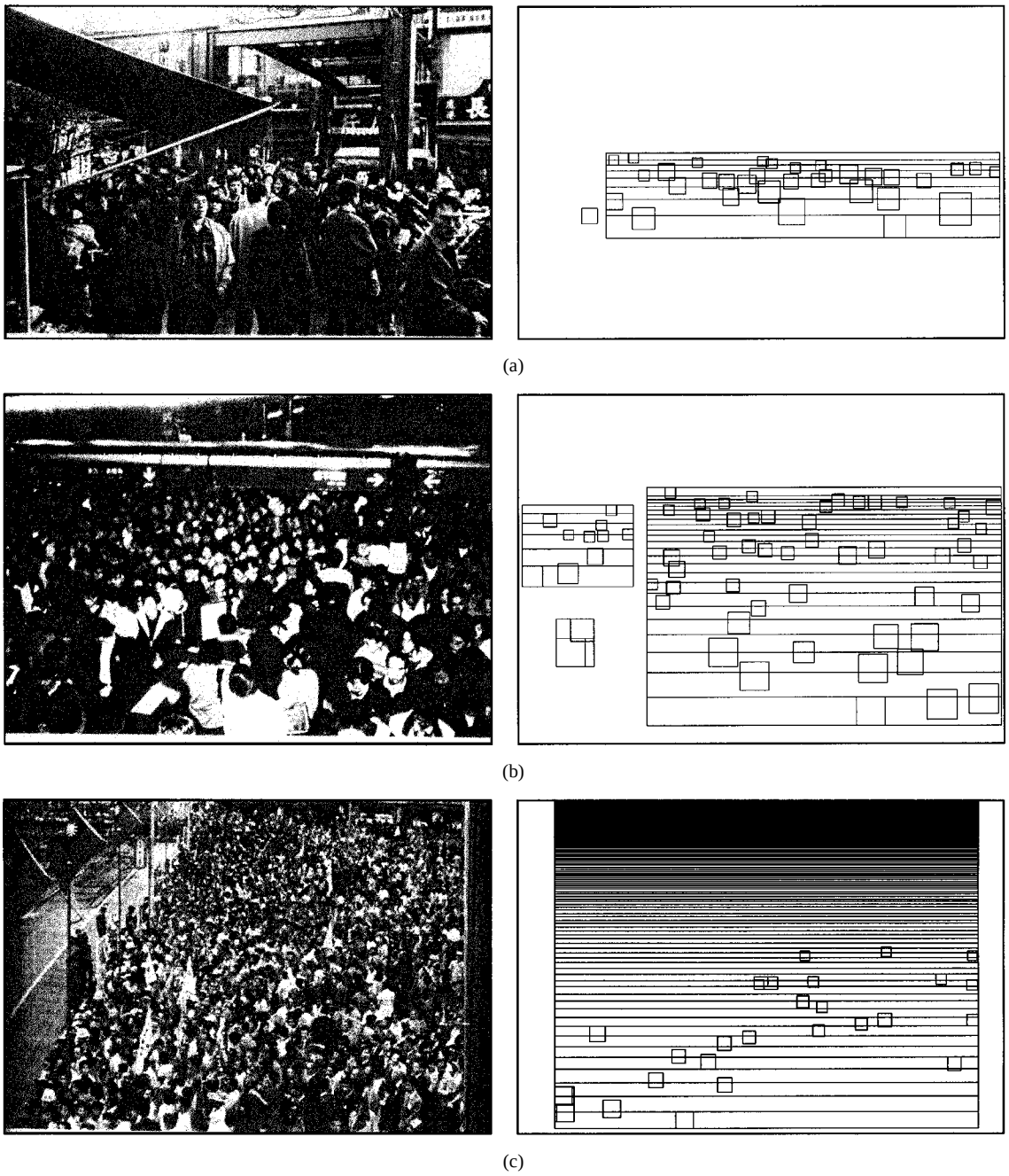


Fig. 13. Some estimating examples in the real world. (a) 88. (b) 222. (c) 1583.

the angle θ_c increases, the estimation accuracy tends to be low. If the angle θ_c is larger, the overlap among much of the crowd is larger such that the effect of estimation is worse. For this reason, the estimation accuracy of the crowd distributed dispersively is higher than the crowd distributed concentratedly. Although the accuracy increases as the visual angle decreases, the number of people in crowds which can be detected also decreases because the field of vision also decreases. To obtain the optimal visual angle which considers both the higher accuracy and crowd size, the optimal value of this angle is 72.5° experimentally.

B. The Results of Estimation of Crowd Size in the Real World

The images that contain different numbers of crowds with simple and complex backgrounds in the real world are tested in this section. All of the images were captured in populous places that were full of different numbers and forms of crowds. In this experiment, the major issue is that the correct number of people in a large crowd cannot be provided to evaluate the performance of estimated results. Therefore, when the number of people in crowds is too large, the system still provides the estimating result.

There was much more noise disturbance in the images captured outdoors than the ones captured in the model world. There are two main influential factors. First, the amount of illumination is influenced by the intensity of light. Even though the image is processed by the histogram equalization first, it still cannot show the clear contours of people if it is too dark or too bright. Second, the textures in a large crowd are complex. As the crowd grows larger, colors and shapes in the crowd become

more complex and the overlap in the crowd is greater. Because all the testing images in the experiment are monochrome images, only the gray level, which refers to a scalar measure of intensity, can be provided without color information. Therefore, even though no one color belongs to the class of head, a wrong detection still will happen if a contour of distribution is circular. As a crowd grows larger and the poses of people change, the circular contours formed by the textures between people increase and the task of estimation is more difficult. As mentioned previously, the overlap in a crowd is the major influential factor, especially in the real world.

Although there are more difficult parts to be detected, the system still can make use of the processed data and the perspective transformation to estimate the approximate number of people in crowds. Fig. 13 shows some estimated examples from the real world. In this paper, the value of the diluted parameter is determined empirically to be 2.3. For each image, the numbers in caption are shown as the estimated number of people in crowds. From the results, it can be seen that the estimating system can work well in the real world even when the noise is serious. Overall, the estimation system is suitable to most conditions in outdoor environments and its performance is good.

IV. Conclusions

In this paper, an approach for the estimation of crowd size by wavelet templates and vision-based techniques is proposed. For calculating the more accurate crowd size in unconstrained environments without a previous reference image or many image sequences, an effective people detector is usually necessary. Generally speaking, only people's heads can be distinguished in a crowd. To describe the characteristics of head, the HWT is used to extract the features of the contour of a head. Using the wavelet template, only significant information that characterizes the contour of a head is evaluated and used. Then, the features are processed by the SVM which is used as a classifier. SVM is an approximate implementation of the method of structural risk minimization, and it has been shown to provide a better general performance than traditional techniques, including NNs. From the detected results processed by these two techniques, the detection performance is shown to be acceptable.

However, even though the detection tool can work well in some cases, it is still limited to some complex situations such as when the overlap among people is great and the contours of the heads are not clear. Therefore, a vision-based technique is proposed to compensate for the lack of detection. By making use of the sizes and positions of the detected frames, information about geometrical projection can be obtained to estimate the approximate number of people in crowds. The proposed approach has a high estimating accuracy rate in experiments.

In the future, the system can be modified along some of the following points to improve the estimating result. As mentioned earlier, the performance of the estimating result is mainly influenced by the kernel factor S which is the diluted parameter and is empirically chosen. If it can be found by a certain theoretical foundation, the estimating result will possibly be nearer to the correct number. On the other hand, if an image sequence can be provided, more information about moving people can be used to further improve the efficacy of head detection. As the number of incorrectly detected or overlooked heads is reduced, the estimated result will become more accurate.

Acknowledgment

The authors would like to thank anonymous referees for their valuable comments and suggestions on revising this paper.

References

[1] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *Electron. Commun. Eng. J.*, vol. 7, pp. 37–47, 1995.

[2] P. Jukkala, J. Ylinan, and A. Raisanen, "Use of a millimeter wave radiometer for detecting pedestrian and bicycle traffic," in *Proc. 17th Eur. Microwave Conf.*, Rome, Italy, 1987, pp. 585–589.

[3] S. S. Mudally, "Novel computer-based infrared pedestrian data-acquisition system," *Electron. Lett.*, vol. 15, pp. 371–372, 1979.

[4] C. S. Regazzoni and A. Tesei, "Distributed data fusion for real-time crowding estimation," *Signal Process.*, vol. 53, pp. 47–63, 1996.

[5] C. S. Regazzoni, A. Tesei, and V. Murino, "A real-time vision system for crowding monitoring," in *Proc. Int. Conf. IECON*, vol. 3, Tokyo, Japan, 1993, pp. 1860–1864.

[6] A. Tesei and C. S. Regazzoni, "Local density evaluation and tracking of multiple objects from complex image sequences," in *Proc. 20th Int. Conf. IECON*, vol. 2, Bologna, Italy, 1994, pp. 744–748.

[7] S. A. Velastin, J. H. Yin, A. C. Davies, M. A. Vicencio-Silva, R. E. Allsop, and A. Penn, "Analysis of crowd movements and densities in built-up environments using image processing," in *Proc. IEE Colloquium Image Processing for Transport Applications*, Aug. 1993.

[8] ——, "Automated measurement of crowd density and motion using image processing," in *Proc. 7th Int. Conf. Road Traffic Monitoring and Control*, London, U.K., 1994, pp. 127–132.

[9] A. N. Marana, L. F. Costa, R. A. Lotufo, and S. A. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. Computer Graphics, Image Processing, and Vision*, Rio de Janeiro, Brazil, 1998, pp. 354–361.

[10] ——, "Estimating crowd density with Minkowski fractal dimension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, Phoenix, AZ, 1999, pp. 3521–3524.

[11] A. N. Marana, S. A. Velastin, L. F. Costa, and R. A. Lotufo, "Estimation of crowd density using image processing," in *Proc. IEE Colloquium Image Processing for Security Applications*, 1997, pp. 11/1–11/8.

[12] ——, "Automatic estimation of crowd density using texture," *Safety Sci.*, vol. 28, pp. 165–175, 1998.

[13] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, pp. 786–804, May 1979.

[14] S.-Y. Cho, T. W. S. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *IEEE Trans. Syst., Man, Cybern. B*, vol. 29, pp. 535–541, 1999.

[15] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.

[16] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proc. 1997 IEEE Comp. Soc. Conf. Computer Vision and Pattern Recognition*, 1997, pp. 193–199.

[17] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Networks*, vol. 10, pp. 1055–1064, Sept. 1999.

[18] A. Polus, J. L. Schofer, and A. Ushpiz, "Pedestrian flow and level of service," *J. Transp. Eng.*, vol. 109, pp. 46–56, 1983.

[19] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1993.

[20] S. Nakamura, *Applied Numerical Methods In C*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

Sheng-Fuu Lin (S'84–M'88) was born in Tainan, R.O.C., in 1954. He received the B.S. and M.S. degrees in mathematics from National Taiwan Normal University in 1976 and 1979, respectively, the M.S. degree in computer science from the University of Maryland, College Park, in 1985, and the Ph.D. degree in electrical engineering from the University of Illinois, Champaign, in 1988.

Since 1988, he has been on the faculty of the Department of Electrical and Control Engineering at National Chiao Tung University, Hsinchu, Taiwan, where he is currently an Associate Professor. His research interests include image processing, image recognition, fuzzy theory, automatic target recognition, and scheduling.

Dr. Lin is a Member of the Chinese Fuzzy System Association and the Chinese Automatic Control Society.

Jaw-Yeh Chen was born in Miaoli, R.O.C., in 1961. He received the B.S. and M.S. degrees in electrical and control engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1984 and 1986, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical and Control Engineering, the National Chiao Tung University. His current research interests are in image processing, image recognition, fuzzy theory, and scheduling.

Hung-Xin Chao was born in Taichung, R.O.C., in 1976. He received the B.S. degrees in automatic control engineering from Fung-Ja University, in 1998, and the M.S. degree in electrical and control engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C. His current research interests are in image processing and image recognition.