Pedestrian Counting System Robust against Illumination Changes

Atsushi Sato Kenji Mase † Akira Tomono Kenichiro Ishii

NTT Human Interface Laboratories †NTT R&D Information and Patent Center
Nippon Telegraph and Telephone Corp.
1-2356 Take, Yokosuka, Kanagawa, 238-03 JAPAN
†1-1-7 Uchisaiwai-cho, Chiyoda-ku, Tokyo, 100 JAPAN

ABSTRACT

This paper presents a realtime pedestrian counting system based on x-t spacetime image analysis. The system counts the number of pedestrians through the combination of three processes: moving object extraction, object counting by region labeling, and direction detection by flow estimation. In order to extract objects reliably under various illumination changes, we propose new moving objects extraction methods that extract moving regions whose shadows have been eliminated by using adaptive background image reconstruction and the color information processing of the images. For labeling and flow estimation, we use *the Ortho-sectioning* method that analyzes extracted regions on a spacetime slice image of the original three dimensional volume. Experimental results confirm the robustness of the system against illumination changes.

1 INTRODUCTION

The volume of pedestrian and automobile traffic must be measured accurately to realize the functions of surveillance, market planning and traffic control. The counting of visitors and customers at convention halls or department stores is frequently performed to assist in the planning of events and business decision making, but it is usually done manually. Conventional manual counting is monotonous labor-intensive work, besides, the results are not reliable. Recently, automatic pedestrian counting systems, which use linear arrays of position sensitive devices(PSDs) such as range sensors, have become commercially available. They count by detecting objects that move across the sensor's field of view. However, there are several problems; the sensors are large, heavy, ineffective in detecting low objects such as children, and cannot capture additional information such as clothing, age and gender. If it were possible to perform automatic pedestrian counting with a TV-camera and image processing technology, the sensor could be very small, easy to set, and it would be possible for the system to count exactly by detecting most objects in the sensing. Then various attributes could also be observed continuously. The system we are now developing realizes these desirable features.

To realize an effective TV camera-based system, we have developed, as a first step, a system with basic functions of directional counting which is equivalent to a PSD based system. The system consists of two main parts: counting and direction detection, and object region extraction adaptive to environmental changes. *The Ortho-sectioning method* in [1] reduces computational time for moving object flow estimation from 3-dimensions to 2-dimensions. The method analyzes extracted regions from a spacetime slice image of the original three dimensional volume. The system employs the method to count pedestrians including detection of moving direction. Section 2 introduces the method and describes how to apply it to binary direction detection. In Section 3, we propose a method to extract moving object regions in a 2-dimensional spacetime domain that is robust against illumination changes. It reliably extracts moving regions whose shadows have been eliminated by using an adaptive background image reconstruction and the color information processing of the images. The method has been implemented on a realtime pedestrian counting system. The processing block diagram and the

0-8194-1369-0/93/$6.00

system implementation of a pedestrian counting system are described in Section 4. The experimental results presented in Section 5 confirm the effectiveness of the system.

2 ORTHO-SECTIONING METHOD

This section introduces *the ortho-sectioning method* which estimates object flow from two orthogonal sectioning plane images of the spacetime volume [1]. The detection method of binary direction is then presented as the one-plane method.

2.1 Ortho-sectioning of spacetime image

If moving objects are viewed with a stationary camera, the spacetime image representation is a 3D volume, in which each object constructs a generalized cylindrical volumetric region within the background as shown in Figure 1. If the sectioning plane, which is a slice of the 3D volume, is skewed against the generator of the cylinder, we call such an image *the non-Epipolar Plane Image*(non-EPI) of the moving object. If the object is non-rigid and the camera angle is adequately set, its non-EPI projection, which can be also called the *Anorthoscopic Projection(anortho-projection)*, reflects much of the object's original shape. Since the anortho-projection of an object has both shape and flow information, we use two orthogonal non-EPI sectioning planes (*ortho-sections*) to get two anortho-projections of each object for flow estimation; this is called the *ortho-sectioning method.*

Several researchers have computed the flow of a moving object on the surface of the generalized cylindrical volumetric region in a 3D spacetime representation from the partial derivatives of the surface and by examining features on the surface. Baker and Bolles introduced generalized *epipolar-plane images*(EPIs) [2]. The intermediate image data produced by arbitrary camera motion could be equated to non-EPI with the generalization. However, they did not utilize the shape information of generalized EPIs, but transformed the images to construct linear EPIs with the help of a known camera path.

The projection of a moving object on the non-EPI plane makes it possible to solve the identification and tracking problems for counting moving objects from an image sequence. Hwang and Takaba[7] built a TV-camera based system that counted pedestrians based on their previous work on an automobile counting system using sonic sensors. This system, however, could count only the number of passing pedestrians, and could not detect their direction of movement. A technique which can discriminate an object's flow is indispensable to construct a complete counting system. Hwang and Takaba's system constructed sparse non-EPIs and utilized the shape information of objects on the plane though, it neglected the flow information contained in non-EPIs.

2.2 Object flow from ortho-sections

A special and efficient ortho-section configuration, comprised of the x-y plane(camera view plane) and x-t *spacetime image* (a spacetime sectioning plane), is used to estimate object flow in this section. Figure 2 shows the ortho-sectioning of an object O moving with flow $\mathbf{V}$. We assume that the normal projection of the object has a primal axis which is skewed by $\theta(0 \leq \theta < \pi)$ against the sampling slit line and that the object's length is H. The movement of the object on the sampling slit line (position is y_s) generates an anortho-projection O' which has the dimensions of width $h(= x_1 - x_0)$ and height $\tau(= t_1 - t_0)$ on the x-t spacetime image. We define the object flow V, and the homogeneous representation by the following equations.

$$\mathbf{V} = \frac{1}{\tau}(h - H\cos\theta, -H\sin\theta), \tag{1}$$

$$\mathbf{V_{ho}} \equiv \mathbf{h} - \mathbf{H} = (h - H\cos\theta, -H\sin\theta, \tau), \tag{2}$$

where $\mathbf{H} \equiv (H\cos\theta, H\sin\theta, 0)$ and $\mathbf{h} \equiv (h, 0, \tau)$ are the primal axes of the normal projection and the anortho-projection of the object, respectively.

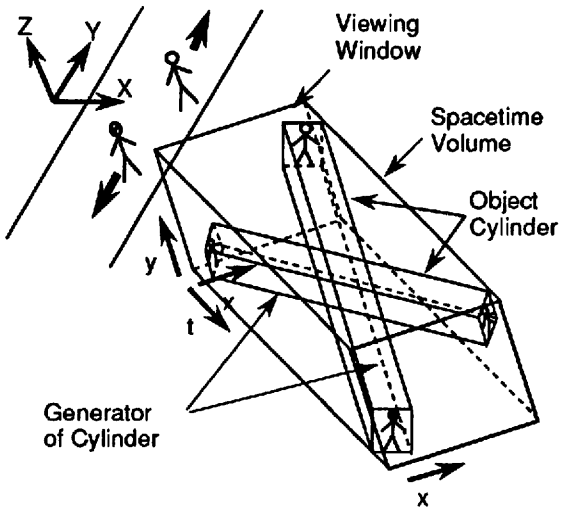

Figure 1: Spacetime Representation of Image Sequence

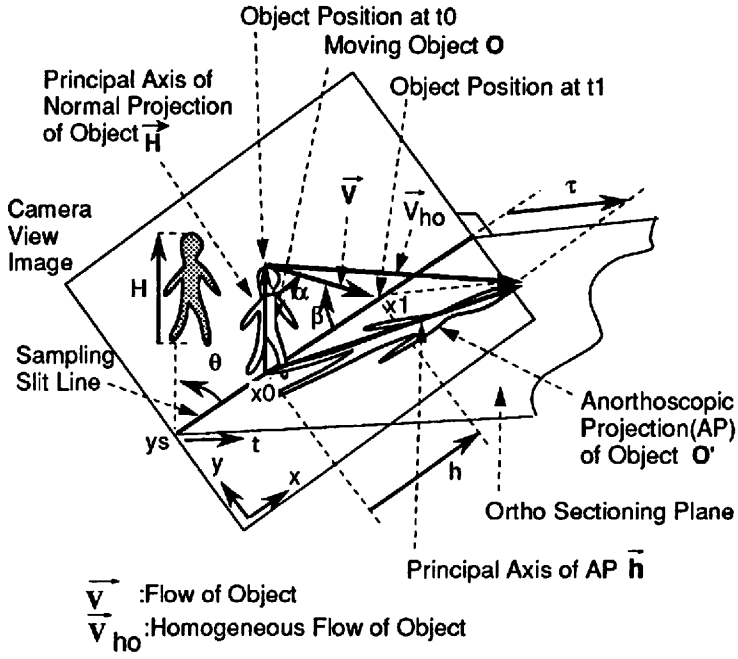

Figure 2: Object Flow by Ortho-Sectioning Method

Object flow is represented by the transposition vector from the primal axis of normal projection **H** to the primal axis of anortho-projection **h**. The dimensions of anortho-projection on x-t spacetime image can be expressed by the following equations which constrain computation possibility and the sampling slit alignment;

$$h = |\mathbf{H}|\frac{\sin\alpha}{\sin\beta}, \tag{3}$$

$$\tau = \frac{|\mathbf{H}|\sin\theta}{|\mathbf{V}|\sin\beta}, \tag{4}$$

where, $\alpha(-\pi < \alpha < \pi)$ is the angle between flow vector **V** and primal axis vector **H** of normal projection and $\beta(0 \leq \beta < \pi)$ is the angle between flow vector and the sampling slit line vector. Here, $\alpha + \beta + \theta = \pi$ (if $0 < \alpha + \theta < \pi$) or 2π(otherwise).

2.3 One sectioning plane method for detecting binary direction

Figure 3 illustrates the relation between the sign of orientation of anortho-projection and the moving direction of the object. We develop the following equation using α, which is the angle of object flow against the primal axis of normal projection;

$$sign(\sin\alpha) = sign(\psi). \tag{5}$$

This can also be confirmed with the following equation derived from equation (3).

$$\sin\alpha = \frac{h}{|\mathbf{H}|}\sin\beta, \tag{6}$$

where $\sin\beta \geq 0$. If $\psi > 0$, $h(= x_1 - x_0) > 0$ and if $\psi < 0$, $h < 0$. Thus, we can estimate moving direction by computing only the signed orientation of the object region against the positive t-axis. This is possible by calculating the second degree moments of each region.

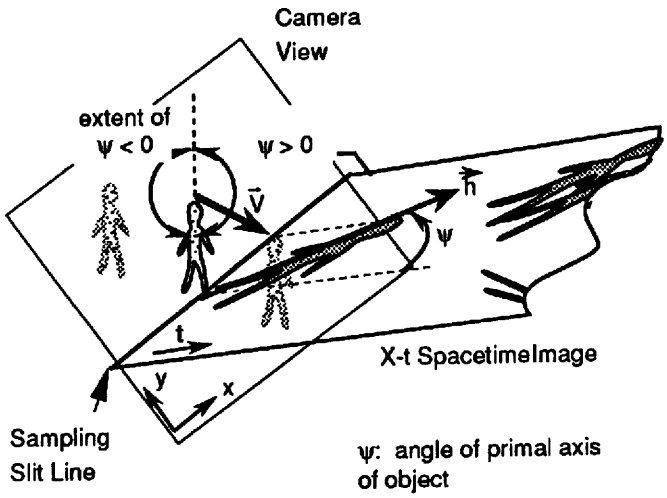

Figure 3: Binary Direction from Anortho-Projection

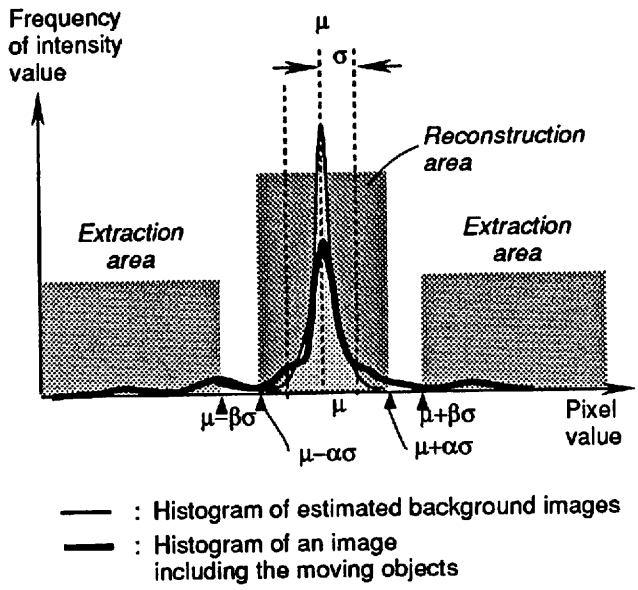

Figure 4: Background Model in Intensity

3 ROBUST MOVING OBJECT EXTRACTION METHOD

This section presents a technique to extract the moving objects from an image sequence by using a background image modeling. The extraction method consists of subtraction and thresholding processes for each pixel on the slit.

In order to extract the moving object from a fixed-angle image sequence, we generally compare the absolute value of the difference between the current frame value and the reference image value and if the absolute value is larger than a certain threshold Th, then it is determined to be a "moving region" by the following equation[3]:

$$d_i(x,y,t) = \begin{cases} 1 & \text{if } \|f(x,y,t_i) - r(x,y)\| > Th \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where, $f()$ is the image value in current frame t_i at spatial coordinates (x,y), $r()$ is the reference image value, $d()$ is the difference image as the binarized result, and Th is the threshold value. As the reference image, we could employ the registered background image if the temporal change of illumination is relatively small against the difference between object values and background values. In order to extract such moving regions stably in a variety of illumination environments, some enhancement techniques of this approach are needed, which include the evaluation and renewal of the background image, and adaptive thresholding.

For background image construction, there are several conventional approaches: for instance, occasional replacement by the current frame image value, adding or subtracting a constant value to or from the currently registered background value, or proportional summation of the current frame image value and the background value[5]. Such techniques use the information in just one frame of the image sequence, however, if the background image could be constructed by using the multi-frame information it is obtained more adaptively to handle comprehensive changes in the illumination environments.

To determine the threshold value adaptively, we generally take the following approach of modeling the image: first, the histogram of the image is approximated by a sum of several functions which present the distribution of image values; second, the threshold value is determined by the relationship among the functions[3][4]. For instance, in the case of the chroma-key-like technique which extracts human images without using a blue

background, the histogram of the whole image plane can be approximated by two functions, which are a normal distribution representing the background variation and a fixed distribution representing the certain values of human image[6]. By searching for the intersection of these functions, the threshold value for extracting objects is estimated. This image modeling technique using the distribution of the plural pixel values is applicable for adaptive determination of the threshold value without being influenced by changes in the scene condition caused by the various illumination changes and the imaging noises.

Shadows cast by objects which change shape in response to changes in the lighting source position, are also difficult to be eliminated in a stable manner. They should be eliminated by employing much more information in the image actively.

Hence, we introduce the following techniques to apply such ideas: 1) the determination of optimum threshold value by adapting to variations in image values, 2) the background reconstruction needed to offset changes in the background image value, and 3) the removal of shadow regions using color information.

3.1 Background modeling for illumination change independent object extraction

In order to extract moving objects in response to illumination changes, we first define the information models of the background for each pixel on the slit using multi-frame images. The background model that reflects the anticipated distribution of the background changes is defined by using the mean and variance of the background pixel values over several frames. Using the parameters of the background model, the threshold for binarization can be appropriately modified to respond to slight background changes caused by imaging noise. For example, in Figure 4, the thick line is the intensity histogram of several scenes at a pixel location, and the thin line is those of corresponding background images.

The background model is composed of three elements for each pixel: the group of the pixel value Y in N frames, which are considered to be grouped as background by a criteria, the average μ and variance σ^2 of these pixel values.

$$\left\{ \begin{array}{l} \{Y_{k-N},,,,,Y_k\} \\ \mu = \frac{1}{N}\Sigma Y_i \\ \sigma^2 = \frac{1}{N}\Sigma(Y_i - \mu)^2 \end{array} \right\} \tag{8}$$

where, Y_i is the intensity value regarding as background image at the frame i.

Using the following equation, the background value is updated and the model is reconstructed, or the object is extracted as a target. If the current pixel value corresponds to the model reconstruction area, the pixel value is regarded as the group of background image value and the model is reconstructed using these values. If the value corresponds to the extraction area, it is regarded as a moving region and is extracted.

$$\left\{ \begin{array}{ll} \textit{Reconstruction}: & \text{if } |x-\mu| < th_r = \sigma \times \alpha(\sigma) \\ \textit{Extraction}: & \text{if } |x-\mu| > th_e = \sigma \times \beta(\sigma) \end{array} \right. \tag{9}$$

where, x is the current pixel intensity value, th_r and th_e are the threshold values, and μ and σ are the average and the standard deviation of the background model respectively. $\alpha()$ and $\beta()$ are the functions of the variation σ, they are limited to avoid the divergence of threshold value through over learning. Thus, this background modeling technique is possible to extract moving object regions with reconstruction of background image in regardless of various illumination changes.

3.2 RGB color space background modeling for shadow removal

Next, we define the background model by using the color information of multi-frame images. A new extraction method proposed is able to extract a moving object from both its shadow and the background.

The color background model is defined by the distribution of the color values over several frames. The distribution is the group of image values of several frames and is represented by several areas in the RGB color

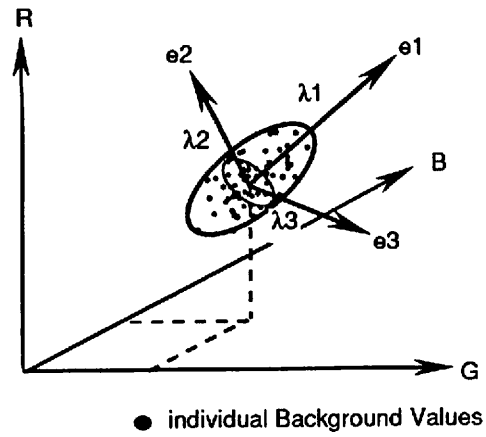

Figure 5: Background Model in RGB Color Space

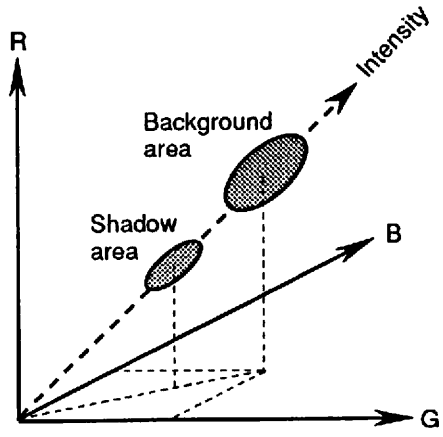

Figure 6: Shadow Model in RGB Color Space

space. The background area, whose distribution is due to the background variation, is approximated as an ellipsoid defined by the principal component values and vectors as shown in Figure 5. Thus, the moving object satisfies the following equation.

$$\{y_1, y_2, y_3\} = \{\ (\vec{x} - \overrightarrow{BG}) \cdot \vec{e}_1, (\vec{x} - \overrightarrow{BG}) \cdot \vec{e}_2, (\vec{x} - \overrightarrow{BG}) \cdot \vec{e}_3\ \} \tag{10}$$

$$\frac{{y_1}^2}{\lambda_1} + \frac{{y_2}^2}{\lambda_2} + \frac{{y_3}^2}{\lambda_3} > Th \tag{11}$$

where, $\vec{x} = (x_r, x_g, x_b)$ is the current image value vector, $\overrightarrow{BG} = (BG_r, BG_g, BG_b)$ is the mean value vector of the background values, $\vec{e_1}, \vec{e_2}, \vec{e_3}$ are the principal component vectors, $\vec{\lambda_1}, \vec{\lambda_2}, \vec{\lambda_3}$ are the principal component values, and Th is the extraction threshold value.

Since the object's shadow could be regarded as the value which lowers the background value along the intensity direction in the RGB color space as shown in Figure 6, the shadow area is defined as the construction of the position and the scale of background model as determined by the following equations:

$$\overrightarrow{BG'} = \gamma \times \overrightarrow{BG}\ ,\quad \vec{\lambda}'_i = \gamma \times \vec{\lambda}_i \quad (\text{i=1,2,3}) \tag{12}$$

$$\{y'_1, y'_2, y'_3\} = \{\ (\vec{x} - \overrightarrow{BG'}) \cdot \vec{e}_1,\ \ (\vec{x} - \overrightarrow{BG'}) \cdot \vec{e}_2,\ \ (\vec{x} - \overrightarrow{BG'}) \cdot \vec{e}_3\ \} \tag{13}$$

$$\frac{{y'_1}^2}{\lambda'_1} + \frac{{y'_2}^2}{\lambda'_2} + \frac{{y'_3}^2}{\lambda'_3} \leq Th' \tag{14}$$

where, γ is the construction coefficient and $0 < \gamma < 1$. Th' is the threshold value for removal of the shadow.

4 PEDESTRIAN COUNTING SYSTEM

This section presents the system configuration of *Pedestrian counting system (the Ped-counter)*. It counts pedestrians and captures their moving direction automatically. The system is built on a graphic workstation with a video frame grabber and no special processing hardware is used. All software is written in the C language.

This system performs processes for 1) extraction from the background, 2) detection of motion, and 3) counting the number of regions. The two methods described earlier in sections 2 and 3 are employed to realize the

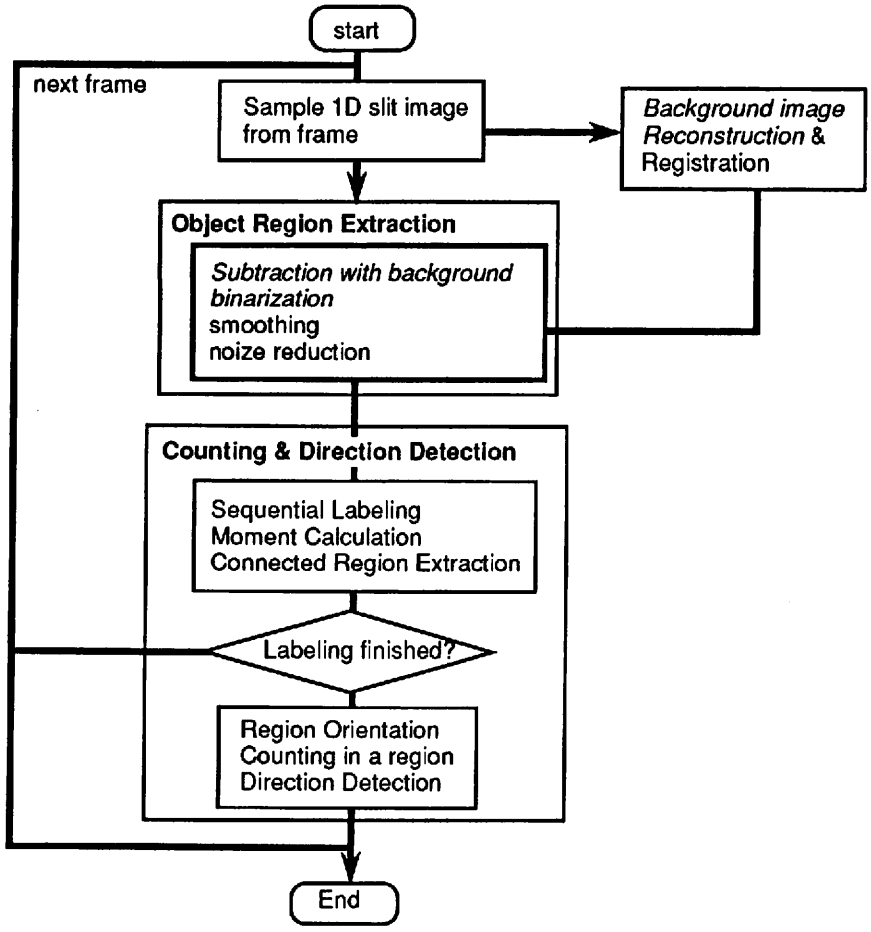

Figure 7: Block diagram of pedestrian counting system

system. The extraction method using the background model extracts moving object regions accurately against various illumination changes, and *the one sectioning plane method* determines the binary moving direction and reduces the amount of computation needed to do so.

Figure 7 illustrates these processes. The system digitizes image sequences from a video input in real time. It then samples the 1D image data on one horizontal sampling slit line and accumulates it sequentially to construct the x-t spacetime image. All pixels on the horizontal line are sampled to construct the x-t spacetime image. RGB color information is used for region extraction. The moving object regions are extracted by eliminating the background image(reference image) at each slit sampling instant. Subtraction of the current frame image from the registered background image gives regions of changing image values. These regions are regarded as extracted objects when they form relatively large regions in the subtracted spacetime image. The registered background image is updated effectively so that the registered background follows illumination changes. The first registration is performed based on a pixelwise histogram calculation after assigning the most significant peak as the background color.

Each extracted region is labeled sequentially, simultaneously with sampling, to find connected component regions. The moments of each connected region are computed and updated in real time. When the scanning of connected regions is finished, the number of pedestrians and their flow in each region are determined. The number of pedestrians is determined by using the size of extracted connected region M_0 as per the following

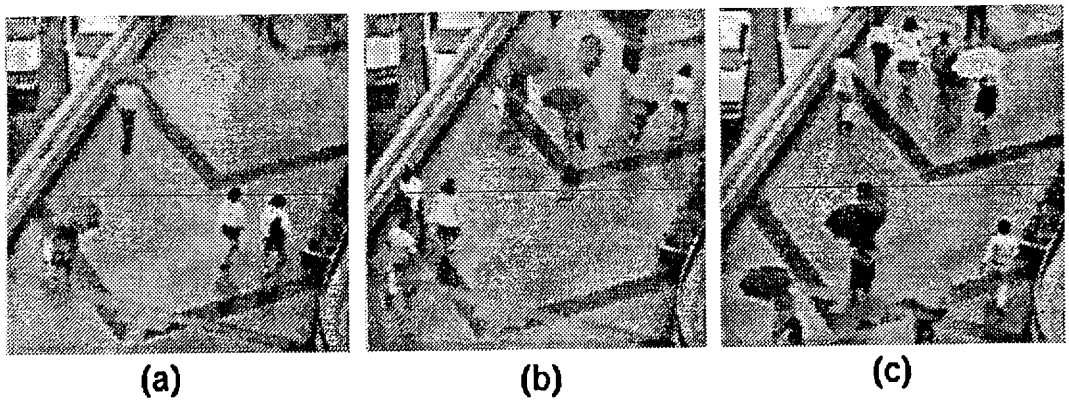

Figure 8: Snapshots of original image sequence

equation.

$$number = \begin{cases} int(M_0/SS) & \text{if } M_0 > T_s \\ 1 & \text{if } T_n < M_0 \leq T_s \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

M_0 is the 0th component of the moment, just the size of the region. If the size is larger than threshold T_s, the region is regarded as overlapped or connected and the area size is divided by a standard area size SS of one person to estimate the number of pedestrians in the group. If the size is smaller than the other threshold T_n, the region is regarded as noise. Next, the *the one sectioning plane method* is used to compute the angle of the primal axis against the positive t-axis, $\psi(-\pi/2 < \psi < \pi/2)$ from the second order of moment. ψ is given as one of the solutions of the following equation;

$$\tan^2\psi + \frac{M_{xx} - M_{tt}}{M_{xt}} \tan\psi - 1 = 0 \tag{16}$$

Ms are components of the moment. The system uses the sign of ψ for direction discrimination.

5 EXPERIMENTAL RESULTS

5.1 Extraction experiment

System performance was evaluated using several test image sequences. The evaluation used several image sequences captured by a video tape recorder and the video tape was set for free run in the experiments. The overall processing cycle included image acquisition and display, slit line data transfer, object extraction, labeling, counting and the display of intermediate data on screen. Total processing cycle time took about 83 ms. Figure 8 are snap shots of the original sequences. The Figures 9(a),10(a) and 11(a) are part of the x-t spacetime image of the image sequences with various conditions; noisy image(Figure 9(a)), illumination change(Figure 10(a)), and shadow casting(Figure 11(a)).

First, we applied the proposed extraction method, which is a background modeling method adapting the illumination change. Figures 9(c) and 10(b) show the processed results. In each image, the line graph indicates the changes in the intensity value at the sampling slit position. The area within both dotted lines corresponds to the model reconstruction area and the remaining areas correspond to the extraction areas. In Figure 9(a), despite variance in the background image due to imaging noise included in the image as a result of imaging conditions, the threshold value for extraction was determined appropriately by adapting the variation. Hence,

Table 1: Result of Counting Experiment(10 min. period, (in/out))

Time(min.)	1	2	3	4	5	6	7	8	9	10	total
ped-counter	29	26	24	36	36	31	28	33	41	26	310
(in/out)	7/22	11/15	6/18	15/21	7/29	18/13	13/15	22/11	16/25	13/13	128/182
actual count	32	25	21	38	35	32	28	34	43	27	317
(in/out)	9/23	12/13	5/16	15/23	6/29	13/19	11/17	18/16	15/28	12/15	116/199

the results in Figure 9(b) show that the human figures were extracted well except for areas missed due to shadows.

In Figure 10, the illumination condition changed greatly as time passed due to clouds. In spite of the varied conditions, the appropriate background value was reconstructed by adapting the change without extracting the moving object region. These results show that the proposed extraction method is able to adaptively determine the threshold value after slight changes and to adjust gradual background changes well.

Next, an experiment was performed to confirm accurate extraction with the color background model. Figure 11(a) shows a typical x-t spacetime image. Figure 11(b) shows the results of binarized by a fixed threshold value using intensity information. In Figure 11(c), the extracted moving objects are colored white, the gray regions represent the shadow, and the black regions show the background. The numbers in Figure 11(c) show the counting results for individual extracted regions. All pedestrians were accompanied by a shadow; nevertheless the extracted regions match the figures of pedestrians much more accurately than the results of the simple intensity binarization. This means that the shadow elimination method can only extract moving objects effectively.

5.2 Counting experiment

Finally, an experiment was carried out to evaluate the counting performance of the pedestrian counting system. The sequence used for the experiment was captured on an outdoor sidewalk, which is similar to the sequence of the example in Figure 11(a), and the illumination conditions were varied. Table 5.1 shows the counting result for a sequence of 10 minutes. For example, 182 people were logged by the system as leaving, but 199 people actually left (128 versus 116 for entering). Overall counting accuracy for the 10 minute period was 98%, which is comparable to the results obtained by Hwang and Takaba[7]. However, considering the illumination conditions in our experiment, which seemed less stable than Hwang and Takaba's experiment, it is clear that our results represent an improvement. On the other hand, the average accuracy for directional counting, which is a new feature of our system, was only 91%. This lower accuracy is due to the estimation error of moving direction which occurs when the regions for pedestrians walking in different directions overlapped. Other counting errors occur when the size of the grouped pedestrian regions becomes smaller with occlusions and when it changes as walking speed changes. The actual accuracy of the range sensor based system and of manual counting in the field has not been published but is said to be around 85%-90%. Overall, the accuracy is comparable to or better than the other counting techniques under the same conditions.

CONCLUSION

We have proposed a method for extracting moving-objects from image sequences using background models. The background models consist of multi-frame information and can extract moving objects while handling background changes and binarizing adaptively. Background modeling in RGB color space is efficient for extracting moving regions without shadows. We implemented the proposed extraction method and the binary detection method for moving direction on a Pedestrian Counting System. Experimental results showed that the pedestrian regions were extracted effectively: the count was accurate and the almost correct walking direction

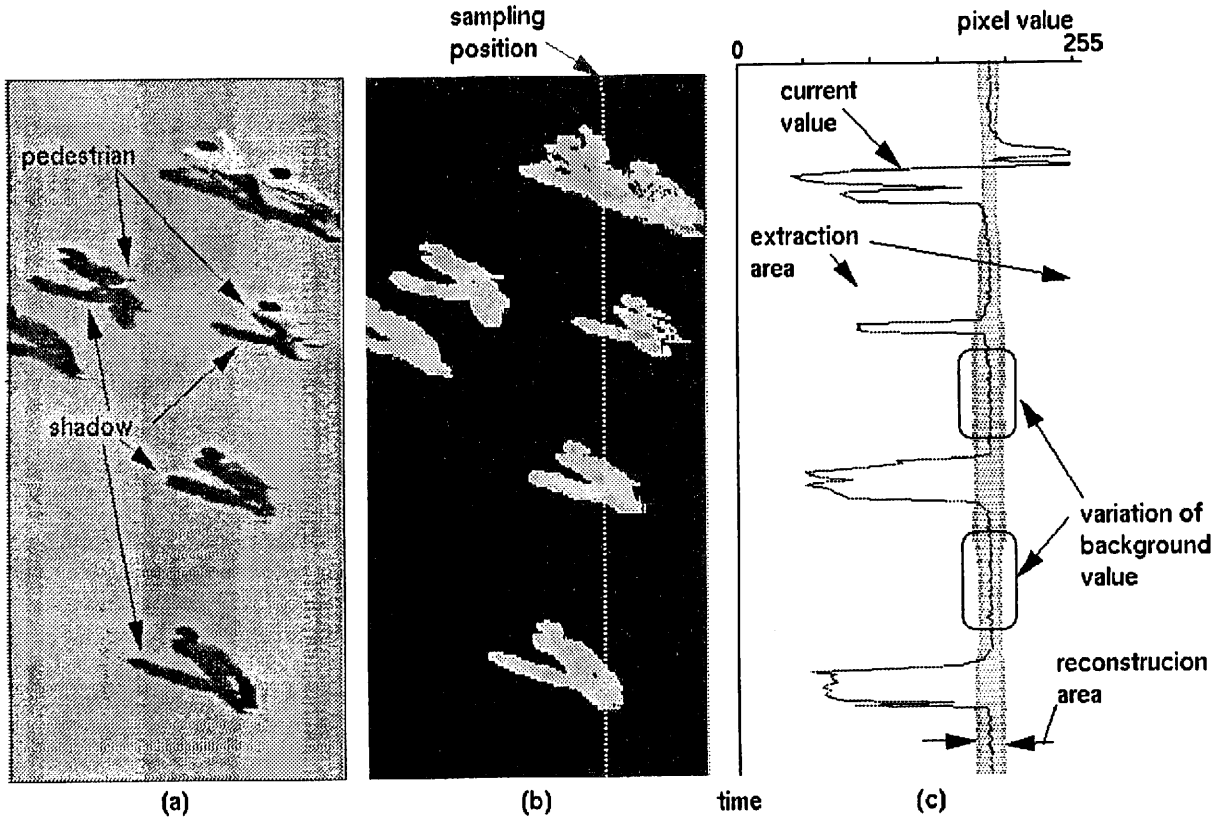

Figure 9: Experimental Results I

was detected. Therefore, this system has sufficient performance for the surveillance of pedestrian traffic as well as events or sales at shopping centers.

ACKNOWLEDGMENT

The authors wish to thank Mr. Kiichi Doi and Mr. Masaki Nitta and for their support to this work. We would also like to thank Dr. Yasuhiko Suenaga and several of our colleagues for their helpful advice and valuable discussions.

REFERENCES

[1] K. Mase, A. Sato, Y. Suenaga and K. Iishii, "A Fast Object Flow Estimation Method based on Spacetime Image Analysis", Proc. IAPR Workshop on Machine Vision Applications '92, pp.199–202,Dec.1992.

[2] H. Baker and R. C. Bolles, "Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface", IJCV,3,pp.33–49.(1989).

[3] R. C. Gonzalez and P. Wintz, "Digital Image Processing (2nd Edition)", Addison-Wesley Publishing Company. Inc., pp354–368,375–382(1987)

[4] A. Rosenfeld and A. C. Kak, "Digital Picture Processing (2nd Edition)", ACADEMIC PRESS, INC., vol 2,pp57–84(1982)

[5] A. Kawabata, S. Tanifuji and Y, Morooka, "An Image Extraction Method for Moving Object", trans. IPS of Japan,Vol.28,No.4,pp.395–402,(1987).(in Japanese).

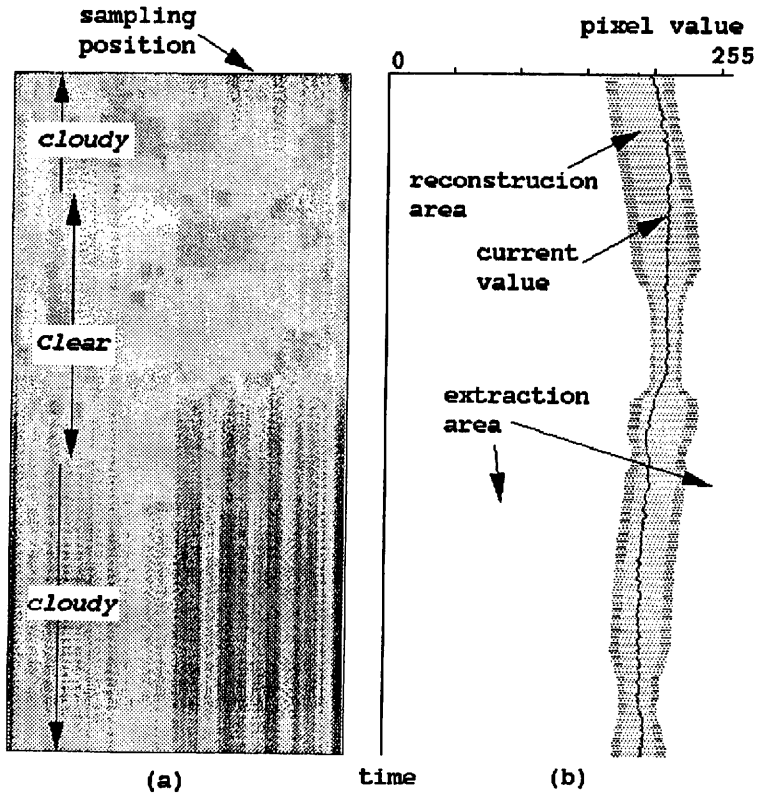

Figure 10: Experimental Results II

[6] T. Fujimoto, M. Shoman and M. Hase, "A Background Region Suppression Method Using Background Memory", Proc. 1991 IEICE Spring Conference,D-452,pp.7-164,(1991).(in Japanese).

[7] B. W. Hwang and S. Takaba, "Real-Time Measurement of Pedestrian Flow Using Processing of ITV images", trans. IEICE of Japan,J66-D,8,pp.917–924,(1983).(in Japanese).

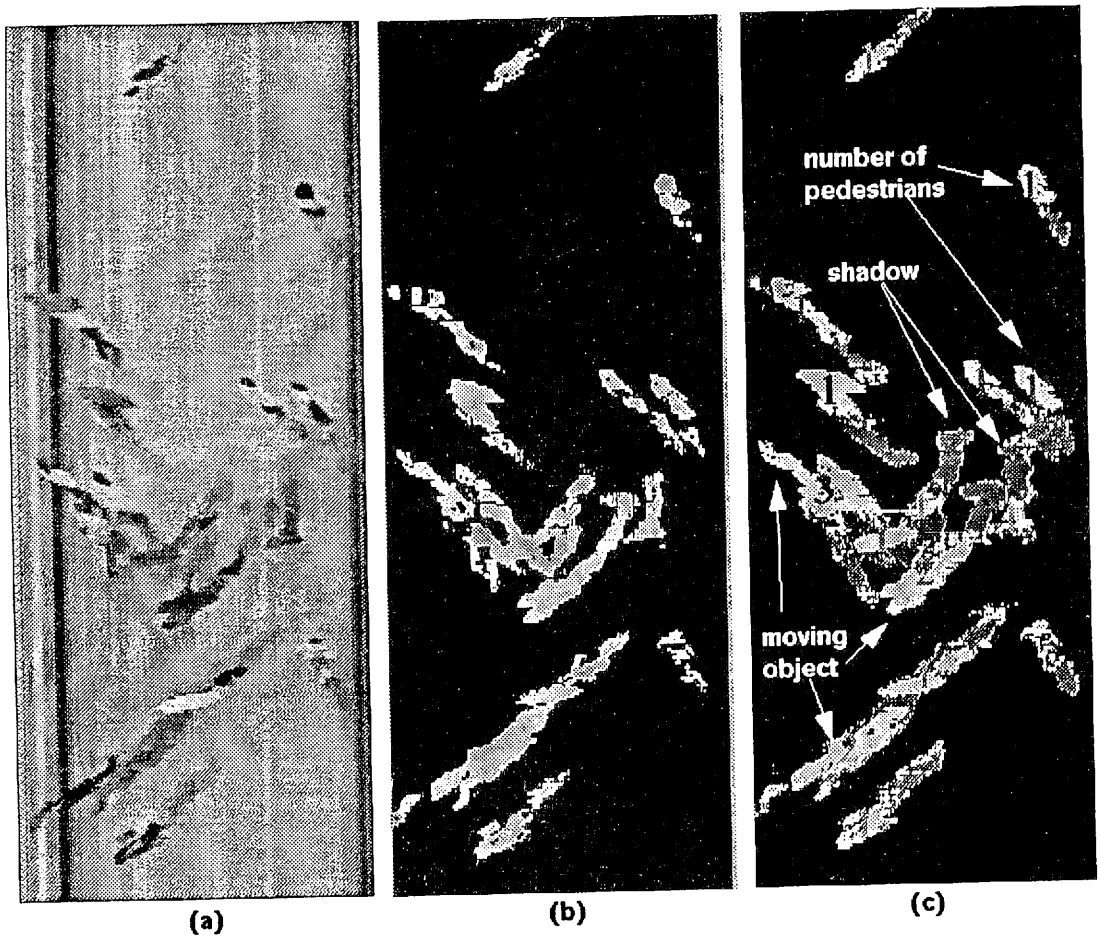

Figure 11: Experimental Results III