Development of a Block-Based Real-Time People Counting System

Hyun Hee Park[1], Hyung Gu Lee[1], Seung-In Noh[2], and Jaihie Kim[1]

[1] Department of Electrical and Electronic Engineering, Yonsei University,
Biometrics Engineering Research Center(BERC),
Republic of Korea
{inextg, lindakim, jhkim}@yonsei.ac.kr
[2] Samsung Electronics, 416, Maetan-3dong,
Yeongtong-gu, Suwon-city, Gyeonggi-do, Republic of Korea

Abstract. In this paper, we propose a block-based real-time people counting system that can be used in various environments including shopping mall entrances, elevators and escalators. The main contributions of this paper are robust background subtraction, the block-based decision method and real-time processing. For robust background subtraction obtained from a number of image sequences, we used a mixture of K Gaussian. The block-based decision method was used to determine the size of the given objects (moving people) in each block. We divided the images into 72 blocks and trained the mean and variance values of the specific objects in each block. This was done in order to provide real-time processing for up to 4 channels. Finally, we analyzed various actions that can occur with moving people in real world environments.

1 Introduction

People counting systems can be used to count or track people, for example at the entrances of shopping malls and buildings. The information can then be used for surveillance purposes, to gather marketing data or to facilitate building management. The use of early automatic counting methods such as light beams, turnstiles and rotary bars led to various problems. These conventional methods could not count people accurately when many individuals passed through the sensors at the same time. To solve this problem, it is necessary for image processing-based approaches to be hance motivated. Thou-Ho et al.[1] presented a bi-directional counting rule, but this method failed in terms of measuring the fixed sizes of objects in the image regions. Terada and Yamaguchi[2] utilized a color camera to extract images of moving people, but the problem of direction-orientation remained intractable. Yoshida et al.[3] used stereo cameras to capture pairs of images, but this method still couldn't solve the problems of counting crowds and direction recognition. The above research[1]-[5] describes how image processing has been used to provide image data that is based on motion analysis, which assumes that the people are moving relative to a static background. In order to focus on dynamic backgrounds, Qi Zang et al.[6] proposed a method of robust background subtraction and maintenance.

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 366–374, 2006.

In this paper, we apply the proposed method to practical and complicated environments such as building gates, escalators, and elevators with a large number of passing people. These practical environments present many problems when compared to limited and simplified environments. To solve these problems, we propose the following methods. First, for robust background subtraction from image sequences, we used a mixture of K Gaussian. Second, we divided images into 6 × 12 sub-blocks and calculated the mean and variance values of the extracted object size of each block. We then plotted these means and variances into a table. Third, we did not use complicated image processing techniques to recognize and track each person. We simply tracked masses of objects. We were able to improve processing time and counting accuracy by using this method. In this paper, we propose a people counting algorithm and present experimental results to verify the effectiveness of the proposed method.

2 Theoretical Approach

2.1 System Configuration

While Figure 1 shows the configuration of our system, Figure 2 shows the overall system block diagram. This is divided into two parts. The first part refers to moving object extraction in image sequences. The second part refers to tracking and counting decisions that are made using the extracted objects. The moving object extraction process consists of four parts. First, we used only the LL part of the Harr wavelet transformed images to remove noise from the input images and down-sample the images ($320 \times 240 \Longrightarrow 160 \times 120$) in order to improve processing time. Second, we produced a reference background model by using a mixture of K Gaussian distributions with N input images. Third, we extracted moving objects by calculating the background subtraction and frame differences between differing time ($t-2$, $t-1$, t) images. Last, we used a morphological mask to remove noise from the images and to fill in the large and small holes existing in the extracted objects. The object tracking and counting decision stage

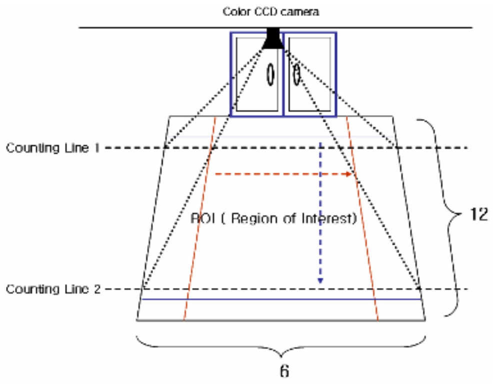


Fig. 1. System configuration

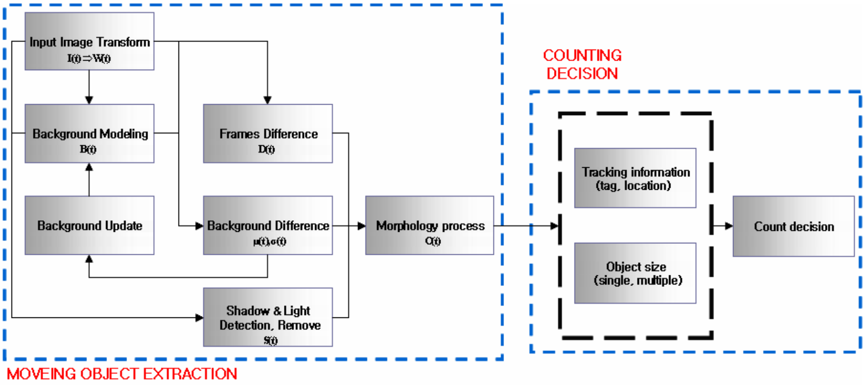


Fig. 2. System overall block diagram

consists of two parts. First, we analyzed the relationship between the extracted objects by using binary images that were obtained from the object extracting stage and compared them with previous ones in order to update correlations among objects and create information when new objects were extracted. Last, we decided to count when objects passed the ROI (region of interest).

2.2 Wavelet Transforms and Background Model

We used the Harr wavelet transform for two reasons; to remove noise from the images by using only a low frequency component, and to down-sample the images by using only the *LL* part. The first and most important step was to extract the moving objects from the background. Each background pixel was modeled by using a mixture of K Gaussian distributions. The K Gaussian distributions were evaluated by using a simple heuristic method to hypothesize which was most likely to be part of the background process. Each pixel was modeled by a mixture of K Gaussian distributions as stated in the formula, where $\mathrm{W_t}$, $\mu_{\mathrm{i,t}}$, K and $\sum_{\mathrm{i,t}}$ are the input images, the mean value of the *ith* distribution, the number of distributions and the *ith* covariance matrix, respectively.

$$\mathrm{P}(\mathrm{W_t}) = \sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{w_{i,t}} * \eta(\mathrm{W_t}, \mu_{\mathrm{i,t}}, \sum \mathrm{_{i,t}}) \tag{1}$$

Previous studies used $K = 3$ for indoor scenes and $K = 5$ for outdoor scenes[6]. So, we made a reference background model by using a mixture of the K Gaussian distributions with the N wavelet transform images. We also made a reference background model with $K = 3$ empirically.

2.3 Shadow and Instant Change Detection

Shadow regions and instant changes of pixels intensity are the main reasons for undesired parts, which affect the final counting results. Therefore, it was necessary to detect and remove these undesired parts. We detected the shadow regions and instant change regions by using the formula below. Once we detected these two regions, we were easily able to remove them to obtain only the desired moving parts.

$$W_{out}(t) = \frac{R_w + G_w + B_w}{3}, B_{out}(t) = \frac{R_B + G_B + B_B}{3} \tag{2}$$

$$S(t) = \frac{W_{out}(t)}{B_{out}(t)} \tag{3}$$

Where $W_{out}(t)$ and $B_{out}(t)$ are the mean values of the R, G, and B components of the present input image and the reference background image, respectively. $S(t)$ is the ratio between $W_{out}(t)$ and $B_{out}(t)$. We were able to decide if a certain area was a shadow region or an instantly changing region by using $S(t)$. Generally, $0 < S(t) < 1$ is defined as a shadow region and $S(t) \geq 1$ is defined as an instantly changing region[7].

2.4 Moving Region Extraction Using Frame Differences

Many errors occur when we extract moving objects only by using background subtraction. To reduce these errors, we used the difference between the frames. The formula below is the procedure used to save $t-2$, $t-1$ and t images in memory and produce the difference images from them. First, we calculated $F(t)$ and $F(t-1)$ using two images. $D(t)$ was calculated from $F(t)$ and $F(t-1)$. $W(t)$ was the Harr wavelet transform image from the original image $I(t)$.

$$F(t-1) = W(t-1) - W(t-2), F(t) = W(t) - W(t-1) \tag{4}$$

$$D(t) = F(t) \vee F(t-1) \tag{5}$$

In this result, we used the extracted moving object combined with the background subtraction result.

2.5 Morphological Process

A morphological process was used to remove noise from the images and fill in small holes of the extracted objects. Practically, an extracted object usually contains large holes that cannot be filled by morphological process. For this problem, we used the masking method instead to fill both the large and small holes effectively and remove noise from the images as well[4]. One disadvantage of this method is the high computational expense. We proposed the modified masking method to reduce the processing time as follows. We produced a proper image to track by using a 5×5 main-mask, which can remove noise from images and fill in holes of the extracted objects. Figure 3 shows this process. This mask will change the pixel value from 1 to 0, if the pixel is determined as noise. Otherwise, it leaves the pixel value as it is. This process sets pixels to the majority pixel value of the mask by counting each pixel value. If a certain pixel value is

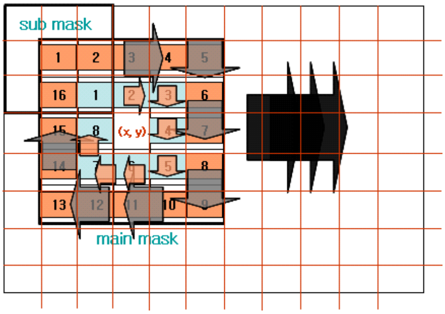


Fig. 3. 5 × 5 main-mask and 3 × 3 sub-mask used for the morphological process

determined to be 0, we move the mask to the next pixel whose value is 1. This process reduces the processing time. If a certain pixel value is determined to be 1, we use a 3 × 3 sub-mask for the surrounding 8 pixels of the center of the 5 × 5 main-mask and count each pixel value to change them 1 → 0, 1 → 1, 0 → 1, 0 → 0. After performing the above steps, we use the same 3 × 3 sub-mask for the other surrounding 16 pixels in the same way. Using this method, we fill in both the large and small holes of the extracted objects and remove the noise from the images at the same time. This would be impossible when using only the morphological process.

2.6 Block-Based Object Counting

The previous method used the same size for the extracted objects of any part of the image and counted objects passing through the interesting area by using this fixed size[1]. This is not an appropriate approach when working in real environments because the size of the extracted objects depends on the position in the image (with a camera at a height of 3.1∼3.3m). So we divided the images into blocks and calculated the mean and variance values of the size of each person for each block. This information was entered into a table. Figure 4 shows the calculating process. The mean and variance values were calculated from the trained images shown in Figure 5. People pass each block at least 10 times in the training images. We trained the images and made a mean and variance table in various environments, such as a 2.8mm lens with a camera at a height of 3.0m, a 2.8mm lens with a camera at a height of 3.1m, and so on. With this table, we were able to conveniently use this system in a wide range of different environments without training.

2.7 Counting Decision Rule

In tracking and counting procedures, it is necessary to analyze merging and splitting relations among people. For example, in light traffic hours, it is simple to count people because they often move independently of one another. However, in busier hours, people frequently merge and split. Therefore, during these busier

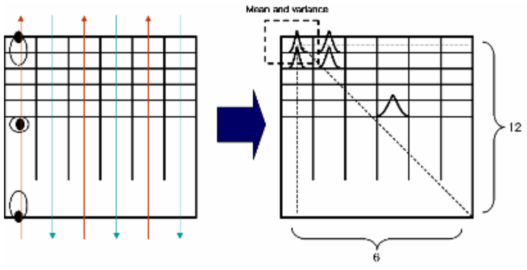
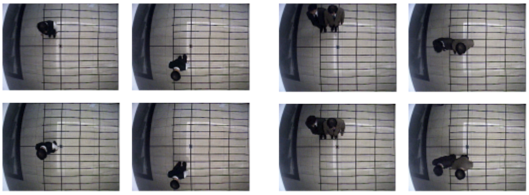
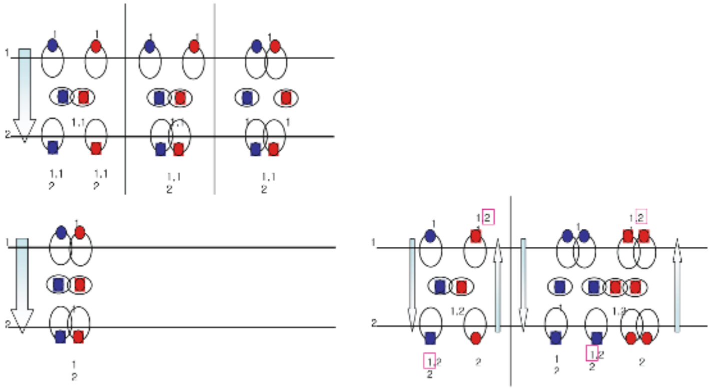


Fig. 4. Divided 6 × 12 blocks and calculation of mean and variance values



(a) (b)

Fig. 5. (a) Images for one-person training image, (b) Images for two-person training images



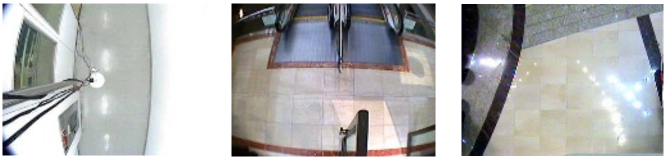
(a) (b)

Fig. 6. (a) Co-directional rule, (b) Bi-directional rule

times it is difficult to count the number of people passing through the ROI and estimate the direction of their paths. To solve this problem, researchers have used specific color information for each person. Although this method is helpful, it requires more processing time, because the system has to store and update all the color information for each person. Hence it is not suitable for real-time systems. It was therefore decided to assign 'tag' information to each moving person, in order to improve time efficiency. This tag information was maintained in the ROI and updated to track the direction of the paths. In this way, no additional image processing or information was needed. Figure 6 shows some examples of the proposed tagging rule in co-directional and bi-directional cases. In Figure 6-(a), when people stepped inside line 1 (the entrance counting line), they were given a label of 'Tag 1'. Similarly, when people stepped outside line 2 (exit counting line), they were given a label of 'Tag 2'. Then, the counting process could be easily performed according to changes of the value of the tag. Also, as shown in Figure 6-(b), the directions of the paths sometimes differ. But if we knew the entrance tag number of each person in advance, we could easily count the passing people by using the alteration of their tag numbers.

3 Experimental Result

We experimented with a camera on the ceiling at a height of H assuming that the height of an average person is h. H and h were measured at approximately 3.1∼3.3m and 150∼180cm, respectively. We used a general CCD camera and a 2.8mm lens. We performed the experiment under practical conditions using three different locations at three different times. Experiments were performed in a corridor, on an escalator, and in an entrance. These locations represented various environments such as light, shadow and highlights caused by strong light. These experimental environments are shown in Figure 7. We used about 20 000 sequential images for each environment. The first 100 frame images were pure background images, which we used to make a reference background model. We produced this model from the first 20 images in the experiments. Table 1 shows the people counting error rates that were recorded for each environment. The TPP (total passing people) parameter represents the counting of the number of people through the ROI (region of interest). The ACE (add counting error) parameter represents values that are higher than the correct count and the UCE (under counting error) parameter represents values that are lower than the correct count (since they exclude children and overweight people). We only trained people of average height and weight. The TCE (total counting error) parameter represents the sum of the ACE and the UCE parameters. Figure 8 shows the people extraction and tracking results in the various environments.



(a) (b) (c)

Fig. 7. Examples of the different environments: (a) Corridor, (b) Escalator, (c) Entrance

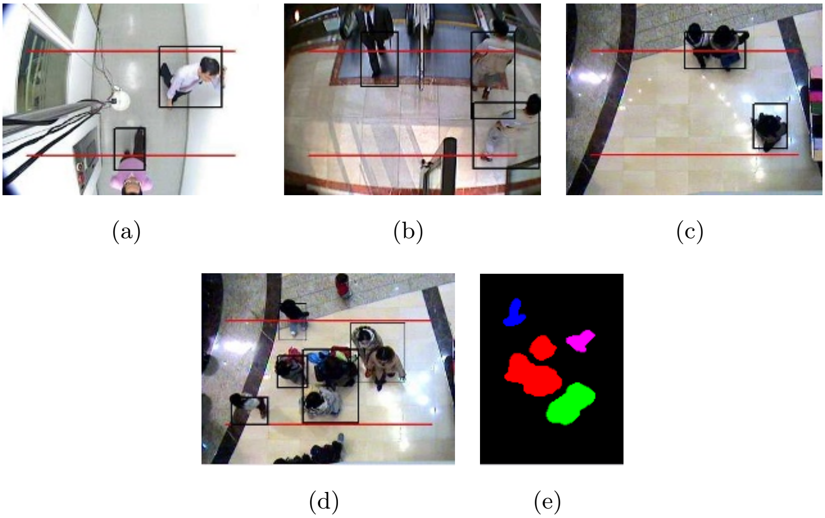


Fig. 8. Results obtained from the different environments: (a) Corridor, (b) Escalator, (c) Entrance, (d) Crowd image, (e) Segment of crowd image

Table 1. Error rates obtained in each environment

Environment	Entrance	Escalator	Corridor
ACE/TPP	10/192(4.16%)	2/207(0.96%)	0/64(0.000%)
UCE/TPP	8/192(5.21%)	6/207(2.89%)	2/64(3.125%)
TCE/TPP	18/192(9.37%)	8/207(3.85%)	2/64(3.125%)

4 Conclusion

In this paper, we proposed a people counting system that can be used to count and track people at entrances, elevators, or escalators where many people are moving. This system is useful for surveillance, building management, and marketing data. We proposed the block-based people counting system which divides an image into 6×12 blocks and trains the size of each person for each block. This proposed method does not detect each person but only tracks masses of objects and counts them by using the trained size of the person. This method improves both accuracy and processing time. We analyzed the time performance with a Pentium 4 3.2 GHz using a video at 320×240 frames (24 bits per pixel) in which our system obtained an average frame rate of 25 fps (performance obtained using the video shown in Figure 8). The counting accuracy was 100% when used with one or two moving people and about 90∼94% when used with three or more moving people.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

References

1. Thou-Ho Chen, “An automatic bi-directional passing-people counting method based on color image processing”, Security Technology, Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on 14-16 Oct. 2003 PP. 200 - 207, 2003.
2. Kenji Terada and Jun’ichi Yamaguchi, “A System for Counting Passing People by Using the Color Camera”, The Transactions of The Institute of Electrical Engineers of Japan.
3. K. Terada, D. Yoshida, S. Oe and J. Yamaguchi, “A Method of Counting the Passing People by Using the Stereo Images”, Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on Volume 2, 24-28, PP. 338 - 342, Oct. 1999.
4. Segen, J., “A camera-based system for tracking people in real time”, Pattern Recognition, 1996., Proceedings of the 13th International Conference on Volume 3, 25-29, PP. 63 - 67, Aug. 1996.
5. Rossi, M., Bozzoli, A., “Tracking and counting moving people”, Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference Volume 3, 13-16, PP. 212 - 216, Nov. 1994.
6. Qi Zang, Klette, R., “Robust background subtraction and maintenance”, Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on Volume 2, 23-26, PP. 90 - 93, Aug. 2004.
7. Hanzi Wang, Suter D., “A re-evaluation of mixture of Gaussian background modeling [video signal processing applications]”,Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP ’05). IEEE International Conference on Volume 2, 18-23, PP. ii/1017 - ii/1020, March 2005.