A Cost-Effective People-Counter for a Crowd of Moving People Based on Two-Stage Segmentation

Chao-Ho Chen¹, Tsong-Yi Chen¹, Da-Jinn Wang² and Tsang-Jie Chen¹

Department of Electronic Engineering¹
National Kaohsiung University of Applied Sciences
No.415, Chien Kung Rd., Kaohsiung 807, Taiwan(R. O. C.)
thouho@kuas.edu.tw

Department of Information Management²
National Kaohsiung Marine University
No.142, Haijhuan Rd., Nanzih District, Kaohsiung 81143, Taiwan(R. O. C.)
wangdaj@mail.nkmu.edu.tw

Received May 2011; revised November 2011

ABSTRACT. This paper is dedicated to a cost-effective people counter for a crowd of moving people by using a zenithal video camera. To obtain a more accurate people count, the two-stage segmentation is developed for extracting each person from a crowd. Firstly, a crowd is segmented by frame-difference technique, followed by morphological processing and region growing. Then, the connected-component labeling method is used to generate many individual people-patterns from the segmented crowd. People-image features, such as the area, height, and width of each people-pattern, are analyzed in order to correctly segment each person from each individual people-pattern. Finally, each person segmented is tracked till touching the base-line and then is counted. Experimental results show that the counting accuracy can be achieved above 91% on average if the crowd moves normally. A comparison with other reported methods of using a zenithal camera manifests the superiority of the proposed method in counting accuracy.

Keywords: People counter, Crowd, Video processing, Moving-object segmentation

1. Introduction. An accurate automatic people counter is very attractive for the entry control and access surveillance of the important military, building security and commercial applications. Anyway, the early automatic people counting systems, such as turn stiles, rotary bar, and light beams, can’t accurately count the people flow when there is more than one person passing through a gate or door at one time. To overcome this problem, many image-processing based approaches with various designs [1]-[14] are hence motivated and they all provide a real-time automatic counting for passing people through a specific region of interest by analyzing a series of images captured with the video camera.

For the transportation applications, Bartolini et al. [1] and Albiol et al. [2] addressed the problems of determining the number of people getting into and out of a bus and train, respectively. To avoid the occlusion problem, Rossi and Bozzoli [3] and Sexton et al. [4] mounted the camera vertically with respect to the floor plane and set the optical axis of the camera in such a way that the passing people could be observed from just overhead. Though, the system [3] based on template motion-estimation tracking may be very time-consuming because the computation complexity increases substantially with the increasing...
number of pedestrians and it may suffer from people-touching overlapping problem. Focused on dynamic backgrounds, Zhang and Sexton [5] developed an automatic pedestrian counting method on an escalator or a moving walkway by using a model-specified directional filter to detect object candidate locations followed by a novel matching process to identify the pedestrian head positions in an image even with complicated contents. With the graylevel-based head analysis, the method will suffer from the following situations: a low contrast of the head image with the background and hair styles or wearing various hats for pedestrians. The first case illustrates that the graylevel technique cannot provide sufficient information for extracting the required pattern from an image, and the second case reveals that a model-based processing may be affected by various sizes and shapes of the human body due to clothing. To increase the count of passing people through a gate at one time, Terada et al. [6] used the stereo images captured by a pair of cameras to cope with both problems of the crowd counting and direction recognition of the passing people. Anyway, the setting of the stereo camera is complicated and the measurement will be seriously sensitive to any shift of camera. To avoid limiting the setting position of the camera and re-counting someone people as they move around, the approach of using multiple cameras located over the region of interest will be an allowable solution [7, 8]. Based on the cost-effective consideration, a single camera with a tracking algorithm may be the better solution and thus Masoud and Papanikolopoulos [9] developed a rectangular model-based recognition of the pedestrian with human motion analysis to achieve a reliable people count. By setting a fixed single camera hung from the ceiling of the gate, Kim et al. [10] proposed a real-time scheme to detect and track the people moving in various directions with a bounding box enclosing each person. Also using a single zenithal camera, Bescos et al. [11] introduced a DCT-based segmentation, which can efficiently consider both lighting and texture information, to cope with some problems, such as shadows, sudden changes in background illumination and sporadic camera motion due to vibration, in order to count people crossing an entrance to a big store. Based on top-view video sequences, a people counting system can be used in different illumination conditions by employing region merging to remove shadows of each object that is segmented by k-means clustering [12]. For counting the passengers getting in/out of a bus, the captured frame by a zenithal camera is firstly divided into many blocks and each block will be classified according to its motion vector. If the accumulated number of blocks with similar motion vectors is greater than a threshold, those blocks are regarded as belonging to the identical passenger pattern. Such a pattern is then judged if it is a passenger to be counted [13]. For solving the frequently-happened overlapping problem [14], called people-image overlapping which is mainly resulted from people touching with each other, both area and color information of people are utilized to count the people flow passing through a gate or door. It also copes with the merge-split problem that people walk sometimes touching with one another and sometimes separating from others. However, the above people counting methods have not presented a solution or cannot provide an accurate count for a crowd of moving people.

In another approach, by taking advantage of human motion analysis, many techniques of the human body tracking or pedestrian detection [15-26] may be applied to the pedestrian counting in open spaces, in which the camera is usually set with a downward-slope view to obscure a more sufficient surveillance range. Nevertheless, the tracking process is always very computation-intensive and their camera setting will make it difficult to segment or recognize each person in the crowded pedestrians owing to a serious overlapping problem. For the purpose of improving the counting accuracy for a crowd of moving people in various illumination conditions, a two-stage-segmentation based people-counting method using a zenithal video camera is proposed. In the proposed method, a crowd is first
segmented by frame-difference technique, followed by morphological processing and region growing. Then, the connected-component labeling method is used to generate many individual people-patterns from the segmented crowd. People-image features, such as the area, height, and width of each people-pattern, are employed in order to correctly segment each person from each individual people-pattern. Finally, each person segmented is tracked till touching the base-line and then is counted. The following section will describe the proposed people-counting algorithm. Then, experimental results, analysis, and comparison are discussed in Section 3 and conclusions are made in the final section.

2. The Proposed People-Counting Algorithm. In the proposed system setting, the camera is set with a downward viewing and hence it has the least affection of people-image overlapping. Figure 1 describes the proposed people counting algorithm which mainly includes crowd segmentation, person segmentation, and person counting and tracking.

![Diagram of the proposed people counting algorithm](image.png)

**Figure 1.** The proposed people counting algorithm.

2.1. Crowd Segmentation. Basically, the background subtraction technique is not suitable for moving-object segmentation if the illumination condition is changeable. The major reason is that the background can be removed completely. Though the optical-flow approach may overcome the above problem, it suffers a computation-intensive problem and thus will not be suitable for the real-time applications. The frame-difference approach can also avoid the above problem, but the extracted moving-object region only appears a rough shape of that object, not the whole body. In the proposed crowd segmentation, the frame-difference technique is first employed to segment the moving-object and then the morphological processing is utilized to obtain a complete shape of each significant moving-object. A region-growing technique is further used to fill the various holes within each moving-object. To refine each body extracted in the above, the color-based body compensation is exploited to derive an optimal moving-object mask. The proposed crowd segmentation process is shown in Figure 2.

As mentioned above, frame-difference will generate a rough hollow body of each pedestrian. To obtain a complete body, dilation-based morphological processing is first used to enhance the boundary and then region-growing is used to fill the holes within each body as possible. Figure 3 describes the process of the region-growing algorithm, where red block denotes a selected seed in the subfigure (b), eight-connection growth (red blocks) is adopted as shown in the subfigure (c), the final result (red blocks) of region-growing is depicted in the subfigure (d). Anyway, from the above region-growing result, there
are still some holes existed in each body. For generating a complete body, a refinement process using color-based compensation is developed in order to further fill those residual holes after executing the region-growing algorithm.

For the frames of region-growing result, let $D_n(x, y)$ denotes the union of differences between the current frame $RF_n(x, y)$ and the previous $k$ frames $RF_{n-j}(x, y)$, $j = 1, 2, \ldots, k$, as illustrated in equation (1). Based on HSV (Hue-Saturation-Value) color model, $DH_n(x, y)$ is defined as hue values of pixels in $D_n(x, y)$, $BH(x, y)$ is defined as hue values of background pixels at the same pixel position $(x, y)$ for those previous $k$ frames, and $DDH_n(x, y)$ is denoted as hue difference between $DH_n(x, y)$ and $BH(x, y)$, as shown in equation (2). If the hue difference $DDH_n(x, y)$ is larger than a color tolerance (CT) value, the pixel of position $(x, y)$ is regarded as foreground pixel (i.e., moving-object pixel). These foreground pixels are then utilized to compensate for those residual holes within the moving-object, as described in equation (3), in which the union of $RF_n(x, y)$ and $OF_n(x, y)$ is the final compensated result $CF_n(x, y)$, where $OF_n(x, y)$ denotes the pixel of original frame.

\[
D_n(x, y) = \bigcup_{j=1}^{k} |RF_n(x, y) - RF_{n-j}(x, y)|
\]

(1)

\[
DDH_n(x, y) = |DH_n(x, y) - BH(x, y)|
\]

(2)

\[
CF_n(x, y) = \begin{cases} 
RF_n(x, y) \cup OF_n(x, y), & DDH_n(x, y) > CT \\
RF_n(x, y), & DDH_n(x, y) \leq CT 
\end{cases}
\]

(3)
TABLE 3. Comparison of count accuracy for five methods using six sequences.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-A-1</td>
<td>In 0 Out 119</td>
<td>0</td>
<td>105</td>
<td>66%</td>
<td>66%</td>
<td>70%</td>
<td>85% 94%</td>
</tr>
<tr>
<td>Video-A-2</td>
<td>78 0</td>
<td>83</td>
<td>0</td>
<td>67%</td>
<td>68%</td>
<td>69%</td>
<td>82% 93%</td>
</tr>
<tr>
<td>Video-A-3</td>
<td>70 74 84 75</td>
<td>68%</td>
<td>71%</td>
<td>72%</td>
<td>83%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>Video-B-1</td>
<td>0 103 0 105</td>
<td>83%</td>
<td>85%</td>
<td>87%</td>
<td>85%</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>Video-B-2</td>
<td>104 0 98 0</td>
<td>65%</td>
<td>67%</td>
<td>67%</td>
<td>82%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>Video-B-3</td>
<td>46 45 57 48</td>
<td>62%</td>
<td>64%</td>
<td>63%</td>
<td>80%</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td>68.5%</td>
<td>70.17%</td>
<td>71.33%</td>
<td>82.83%</td>
<td>91.67%</td>
</tr>
</tbody>
</table>

Acknowledgment. This work is supported by National Science Council under Grant NSC 99-2622-E-151-015-CC3, Taiwan, R.O.C.

REFERENCES


