K-MEANS BASED SEGMENTATION FOR REAL-TIME ZENITHAL PEOPLE COUNTING

Borislav Antić, Dragan Letić, Dubravko Ćulibrk and Vladimir Crnojević

Faculty of Technical Sciences, University of Novi Sad, Serbia

ABSTRACT

The paper presents an efficient and reliable approach to automatic people segmentation, tracking and counting, designed for a system with an overhead mounted (zenithal) camera. Upon the initial block-wise background subtraction, k-means clustering is used to enable the segmentation of single persons in the scene. The number of people in the scene is estimated as the maximal number of clusters with acceptable inter-cluster separation. Tracking of segmented people is addressed as a problem of dynamic cluster assignment between two consecutive frames and it is solved in a greedy fashion. Systems for people counting are applied to people surveillance and management and lately within the ambient intelligence solutions. Experimental results suggest that the proposed method is able to achieve very good results in terms of counting accuracy and execution speed.

Index Terms— People counting, People segmentation, Background subtraction, People tracking.

1. INTRODUCTION

Efficient and reliable automatic people detection, tracking and counting is of interest in a number of commercial applications. People surveillance at certain public places such as underground stations, football pitches, and public buildings rely on such systems with important implications to human safety. In addition to improving security, information obtained from such systems can be used to identify hourly traffic patterns, optimize labor scheduling, monitor the effectiveness of promotional events, etc.

Apart from image sensors, contact and other sensor technologies are used for people counting [1]. Systems using contact-type counters, such as turnstiles, count people only one at a time and can obstruct the passage way and cause congestion if there is a high-density traffic. Due to their design, they are prone to undercounting. Systems using infrared beams or heat sensors do not block the doorways, but also do not provide enough accuracy to distinguish people in a group. Clearly, a more informative type of sensors is needed, and cameras are certainly a reasonable choice [2] [3].

Foreground segmentation is the first step in many computer vision applications. The segmentation of foreground regions for human detection is usually obtained by calculating the pixel-wise differences between the current frame and the background image, followed by an automatic thresholding [1] [4] [5]. If the additional accuracy provided by pixel-wise approaches is not warranted [6] [7], block-wise approaches are more preferable since they produce more stable segmentation in the presence of changing light or shadow effects.

Rossi and Bozzoli [8] use gray level features sensitive to high frequency changes in the scene to detect moving objects. They use template matching to track the extracted features. Huang and Chow [6] use more elaborate features describing the shape of the foreground blobs, which they dub datacurves. Rather than tracking they simply count the number of persons in the region of interest. Velipasalar *et al.* [1] propose the use of the size of detected blobs to segment single persons and mean-shift procedure as a way to handle the merging of blobs. In [9], Beleznai *et al.* also employ the mean-shift procedure to develop a general people tracking system.

The overhead position of the camera effectively removes the object occlusion problem that otherwise has to be handled in a more general setup [9] [10]. However, the segmentation results often contain merged blobs pertinent to several people standing close to one another. In [5], Snidaro *et al.* develope a people counting system designed for the ambient intelligence (AmI) applications, that assumes the top-view placement of cameras. An extended set of features is used to describe bounding boxes pertinent to the detected foreground blobs, and Kalman and Mean-shift filters track these bounding boxes.

Barandiaran *et al.* [11] have recently described a real-time people counting system, that uses a single overhead mounted camera, and achieves counting accuracy of 95%. Counting is performed inside the image zone with multiple counting lines placed over it.

An efficient approach to people segmentation, tracking and counting is presented in this paper. The method is designed for a zenithal camera system similar to those in [1] [5] [6] [8]. The presented method is simple, robust and achieves good results.

The organization of the paper is as follows. Section 2 describes the method for people counting in detail. Experimental validation is presented in Section 3. Conclusions and some directions for future work can be found in Section 4.

This research activity has been supported by EUREKA Project.

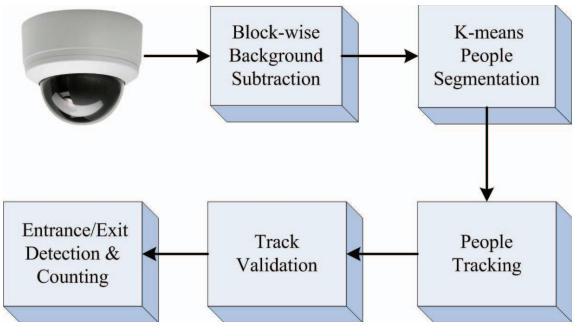


Fig. 1. Block diagram of the proposed system for people counting.

2. SYSTEM ARCHITECTURE

The block diagram of the proposed system for people segmentation, tracking and counting is given in Fig. 1. The system uses a zenithal video camera as a standard setup for people counting system. This type of camera provides the best possible view of the scene in terms of people detection and tracking, as it minimizes the occlusion and the objects appear relatively constant in size. To ensure lightweight operation, necessary for applications in embedded systems, the proposed algorithm is block-wise, which significantly reduces the amount of computation.

The first step of the algorithm is background subtraction, that is in our setup performed on a block level. It detects those image blocks that belong to the foreground by comparing image blocks from the current frame with those from the background image. Background image is obtained through a recursive filtering

$$BG^{t+1}_{m,n,p}(i,j) = (1-\alpha)\cdot BG^{t}_{m,n,p}(i,j) + \alpha \cdot I^{t}_{m,n,p}(i,j) \quad (1)$$

Indices (m,n) refer to the block coordinates, p to image channel (Red, Green, Blue), t denotes the frame number, indices (i,j) refer to pixel coordinates within the block, and α is the learning rate (typical values are $0.01-0.1$). Standard block sizes are 8×8, 12×12 and 16×16.

In the case of illumination changes or shadows, all three image channels are multiplied by approximately the same constants within one block. The following model is assumed for blocks in the background

$$I^{t}_{m,n,p}(i,j) = \beta_{m,n,p}\cdot BG^{t}_{m,n,p}(i,j) + W^{t}_{m,n,p}(i,j), \quad (2)$$

where W denotes additive white Gaussian noise (AWGN), and $\beta_{m,n,1} \approx \beta_{m,n,2} \approx \beta_{m,n,3}$.

Multiplicative factors $\beta_{m,n,p}$ are determined using maximum likelihood estimation (MLE)

$$\beta_{m,n,p} = \frac{\sum_{i,j} I^{t}_{m,n,p}(i,j)\cdot BG^{t}_{m,n,p}(i,j)}{\sum_{i,j}\left(BG^{t}_{m,n,p}(i,j)\right)^2} \quad (3)$$

Multiplicative factors of blocks that belong to background have approximately equal values close to 1. The detection of foreground blocks is based on a divergence measure of multiplicative factors. In this paper, difference of multiplicative factor's maxima and minima is used as divergence measure.

$$\delta\beta_{m,n} = \max_{p}\beta_{m,n,p} - \min_{p}\beta_{m,n,p} \quad (4)$$

If the divergence measure is not small or some multiplicative factor deviates significantly from 1, the block is denoted as foreground.

$$FG_{m,n} = \begin{cases} 1, & \text{if} \quad \delta\beta_{m,n} > T_1 \vee |\beta_{m,n,p} - 1| > T_2 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

We found that the following threshold values give satisfactory results: $T_1 \in [0.1, 0.2]$ and $T_2 \in [0.3, 0.6]$.

People segmentation is a difficult problem in image analysis. In the zenithal camera setup, the problem is somewhat relaxed. Observed from above, people are seen as concentrated shapes that can be extracted using standard clustering techniques such as k-means. However, k-means algorithm assumes that the number of clusters is known a priori. Therefore, a value of k is estimated as the maximal number of clusters for which the inter-cluster distance is still above some minimal allowable inter-cluster distance D_{min}. This constant corresponds to the average person size, that can be established through experiments. If there are k clusters in the image, with their centroids $C_i,\ i = 1,2,...,k$, the minimal inter-cluster distance is defined as

$$d^{k}_{min} = \min_{1\le i<j\le k}\|C_i - C_j\| \quad (6)$$

If there is only one cluster, we formally define $d^{1}_{min} = \infty$. Actual number of clusters k^* is estimated as the maximal number of clusters that still has the minimal inter-cluster distance d^{k}_{min} higher than D_{min}.

$$k^* = \max\{k | d^{k}_{min} \ge D_{min} \wedge d^{k+1}_{min} < D_{min}\} \quad (7)$$

People tracking is based on a greedy solution to the dynamic assignment problem between the clusters in two consecutive frames. The aim is to minimize the total Euclidean squared distance between corresponding clusters. Greedy algorithm finds two clusters that are at the minimum distance. If they are close enough, the algorithm relates them and tries to find next cluster assignment for the remaining clusters.

Two auxiliary lines have been entered to demarcate the counting zone. Only tracks spanning more than a half of the

(a) (b) (c)

Fig. 2. Demonstrative results of the processing steps for background subtraction, people segmentation, tracking and counting. (a) Original frame with the results of people tracking and counting shown on it (red lines denote the borders of counting zone). (b) Block-wise background subtraction. (c) K-means people segmentation.

counting zone are regarded as valid. The algorithm checks the end points of each validated track, and finds the counting zone's border lines lying closest to them. If these lines are opposite to one another, the entrance (or exit) counter is incremented.

Fig. 2 demonstrates the main steps of the algorithm. Left picture shows the original frame with the results of people tracking and counting on it. Center picture illustrates the proposed block-wise background subtraction algorithm (block size is 8×8 pixels). Right picture depicts the result of k-means based people segmentation, where the number of clusters is automatically determined using minimal inter-cluster distance.

3. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the performance of the proposed algorithm, two indoor color video sequences were generated by zenithal cameras mounted $3m$ above the ground. The videos were captured in an office (640×480 pixels, 1635 frames, fluorescent light) and in a corridor (320×240 pixels, 1500 frames, daylight). Both sequences contain at most three persons in the scene at the same time.

We compared the proposed algorithm with recently published method for people counting by Barandiaran *et al.* [11]. The result of this comparison is given in Table 1. Abbreviations TP, FP and FN designate the numbers of true positives, false positives and false negatives, respectively. Precision, recall and F-score are defined as follows

$$precision = \frac{TP}{TP + FP} \tag{8a}$$

$$recall = \frac{TP}{TP + FN} \tag{8b}$$

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{8c}$$

Table 1. Comparison of methods for people counting.

	Ground Truth	Barandiaran *et al.* [11]	Proposed method
in+out	11 + 10	11 + 9	10 + 10
TP	21	19	20
FP+FN	0 + 0	1 + 2	0 + 1
precision	1.00	0.95	1.00
recall	1.00	0.90	0.95
F-score	1.00	0.92	0.97

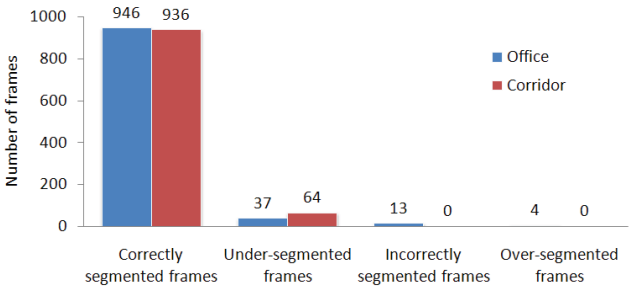
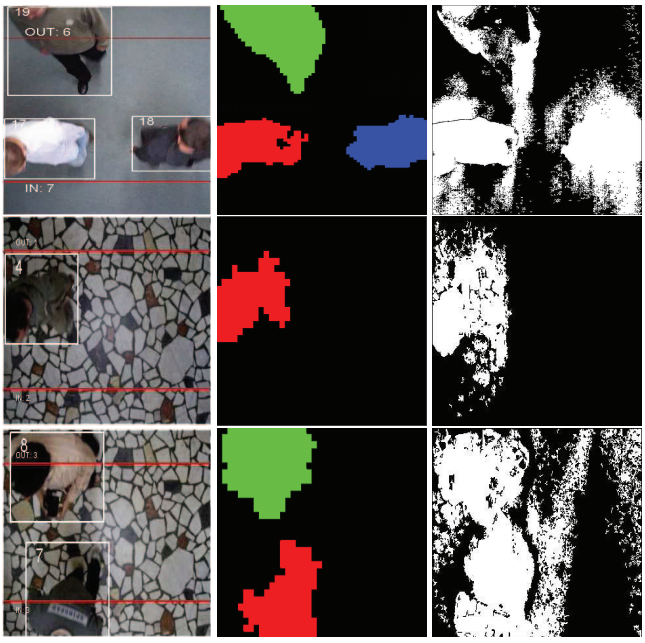


Fig. 3. Statistical analysis of the k-means based segmentation for non-empty frames of video sequences *Office* and *Corridor*. Vertical bars show the number of frames that were correctly segmented, under-segmented (more persons than clusters), over-segmented (more clusters than persons) and incorrectly segmented (the same number of clusters and persons, but the clusters do not correspond well to single persons). All numbers are relative to 1,000 frames.

Statistical evaluation of the proposed segmentation procedure is given in Fig. 3. Although simple and straightforward, the algorithm for people segmentation shows very good performance. The correct segmentation rate of almost 95% of all non-empty frames significantly contributes to high precision and recall values of the counting process yielding the F-score of 0.97.

Fig. 4 shows the results of people segmentation, tracking and counting and also the comparison of methods for background subtraction. Fig. 4(a) illustrates the performance of people tracker, which appropriately assigns the clusters from previous frame to the clusters in current frame, if the frame is correctly segmented. Results of the proposed block-wise background subtraction and k-means based people segmentation are given in Fig. 4(b). Background subtraction results according to the Gaussian Mixture Model (GMM) [12] are given in Fig. 4(c) for comparison. Block-wise background subtraction is clearly more resilient to shadows.

The proposed algorithm was originally implemented in MATLAB, and subsequent implementation in C++ demonstrated real-time performance (12 fps, 3.0 GHz PC). Related video materials can be found at `www.ursusgroup.com`.



(a) (b) (c)

Fig. 4. The results of people segmentation, tracking and counting and the comparison of methods for background subtraction. (a) People tracking and counting results for sample frames from *Office* and *Corridor* video sequences. (b) Proposed block-wise background subtraction and people segmentation. (c) Background subtraction according to the Gaussian Mixture Model (GMM) [12].

4. CONCLUSION

A novel method for people segmentation, tracking and counting in a zenithal camera system is presented in the paper. The algorithm uses a block-wise background subtraction and relies solely on the k-means based clustering of foreground blocks to achieve the segmentation of humans. Greedy solution to the problem of dynamic cluster assignment between two consecutive frames is exploited for people tracking. Representative video sequences have been used to evaluate the result. The proposed people counting method is simple, yet efficient, and achieves real time performance. The performance of the algorithm could be further improved by using multi-dimensional dynamic cluster assignment, mean-shift tracking and simultaneous tracking and segmentation.

5. REFERENCES

[1] S. Velipasalar, Y.-L. Tian, and A. Hampapur, "Automatic counting of interacting people by using a single uncalibrated camera," *Multimedia and Expo, IEEE International Conference on*, vol. 0, pp. 1265–1268, 2006.

[2] A. J. Schofield, P. A. Mehta, and T. John Stonham, "A system for counting people in video images using neural networks to identify the background scene," *Pattern Recognition*, vol. 29, no. 8, pp. 1421–1428, 1996.

[3] C. Sacchi, G. Gera, L. Marcenaro, and C. S. Regazzoni, "Advanced image-processing tools for counting people in tourist site-monitoring applications," *Signal Process.*, vol. 81, no. 5, pp. 1017–1040, 2001.

[4] Shao-Yi Chien, Yu-Wen Huang, Bing-Yu Hsieh, Shyh-Yih Ma, and Liang-Gee Chen, "Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques," *IEEE Transactions on Multimedia*, vol. 6, no. 5, pp. 732–748, 2004.

[5] L. Snidaro, C. Micheloni, and C. Chiavedale, "Video security for ambient intelligence," *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, vol. 35, no. 1, pp. 133–144, 2005.

[6] D. Huang and Tommy W. S. Chow, "A people-counting system using a hybrid rbf neural network," *Neural Process. Lett.*, vol. 18, no. 2, pp. 97–113, 2003.

[7] J. Bescos, J.M. Menendez, and N. Garcia, "Dct based segmentation applied to a scalable zenithal people counter," in *Proc. of the IEEE International Conference on Image Processing*, Sept. 2003, vol. 3, pp. III–1005–8 vol.2.

[8] M. Rossi and A. Bozzoli, "Tracking and counting moving people," in *Proc. of the IEEE International Conference on Image Processing*, Nov 1994, vol. 3, pp. 212–216 vol.3.

[9] C. Beleznai, B. Frühstück, and H. Bischof, "Human tracking by fast mean shift mode seeking," *Journal of Multimedia*, vol. 1, no. 1, pp. 1–8, 2006.

[10] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 279–289, 2006.

[11] J. Barandiaran, B. Murguia, and F. Boto, "Real-time people counting using multiple lines," in *WIAMIS '08: Proc. of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, Washington, DC, USA, 2008, pp. 159–162, IEEE Computer Society.

[12] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1999, vol. 2, pp. 246–252 Vol. 2.