A Real-Time People Counting Approach in Indoor Environment

Jun Luo[1], Jinqiao Wang[2], Huazhong Xu[1], and Hanqing Lu[2]

[1] School of Automation, Wuhan University of Technology, Wuhan, China
[2] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
junjing2218@gmail.com, wutxhz@163.com, {jqwang,uhq}@nlpr.ia.ac.cn

Abstract. Due to complex background information, shadow and occlusions, it is difficult to count people accurately. In this paper, we propose a fast and robust human counting approach in indoor space. Firstly, we use foreground object extraction to remove background information. In order to get both moving people and stationary people, we designed a block-updating way to update the background model. Secondly, we train a multi-view head-shoulder model to find candidate people, and an improved k-means clustering is proposed to locate the position of each people. Finally, a temporal filter with frame-difference is used to refine the counting results and detect noise, such as double-count, random disturbance. An indoor people dataset is recorded in the classroom of our university. Experiments and comparison show the promise of the proposed approach.

Keywords: People counting, block-updating, improved k-means clustering, temporal refinement.

1 Introduction

Counting the number of human indoors is a challenging problem that has lots of practical applications, such as building security, room resources adjustment, market research, and intelligent building etc. In indoor environments, as a moving pixel extraction, the traditional background subtraction method is limited because not all humans' bodies are moving. What's more, when people get together, we will get a large blob with several objects inside. These blobs can not provide the object level information and are hard to segment. Actually, most of the time the number of human remain stable in indoor spaces. In other words, Although occlusions often occur, the room is at a dynamic stability state. Therefore, people counting in indoor environment is a challenging topic.

2 Related Work

A direct top-down view can avoid most occlusions in people counting, Teixeira and Savvides [5] proposed a lightweight method for localizing and counting people in indoor spaces using custom-built camera installed on the ceiling. Li et al. [2]

X. He et al. (Eds.): MMM 2015, Part I, LNCS 8935, pp. 214–223, 2015.
© Springer International Publishing Switzerland 2015

improved the accuracy rate of people counting by analyzing across multiple cameras. The temporal continuity of objects is an important reason to achieve stable pedestrian counting. Zhao and Nevatia [8] treated problem of segmenting individual humans in crowded as a model-based Bayesian segmentation problem. They presented an efficient Markov chain Monte Carlo (MCMC) method to get the solution. Wang et al. [6] built a spatio-temporal group context model to model the spatio-temporal relationships between groups, formulate the problem of pedestrian counting as a joint maximum a posteriori (MAP) problem. Zhang and Chen [7] used group tracking to compensate weakness of multiple human segmentation, which can handle complete occlusion. Chan and Vasconcelos [1] using dynamic texture, segmented the scene into different regions with different motions, extracted various features from each segment. A Gaussian process is used for estimating the pedestrian count for each segment. Most of the previous works focused on how to get an accurate counting result at the expense of real-time counting. Moreover, different from counting outdoors, the assumption that people would be in motion for a substantial amount of time is invalid, the number of human indoor changes little during most periods. The key problem of counting indoor is how to get a stable and accurate number of human in these periods.

3 Proposed Framework

As shown in Fig.1, our people counting method mainly includes three modules: foreground object extraction, head-shoulder detection and temporal refinement. Firstly, through the background subtraction, we will get blobs and corresponding gradient maps with human bodies inside. Then, a multi-view head-shoulder model is trained for head-shoulder detection. A head detection and a head-shoulder detection can find the general position of human, vice versa, results of head-shoulder detection will help update the background model with a block-updating method. After that, the clustering method is used for obtaining the counting results. Finally, the frame-difference method is utilized to estimate occlusions and refine the counting result.

3.1 Foreground Object Extraction

Traditional background subtraction approaches segment video frames into stationary pixels and moving pixels. The background model is built based on stationary pixels as a reference of moving pixels. Similarly, we also can consider that frames are constituted by pixels on and off human bodies. We can build and update the background model with no-human pixels to extract pixels on human bodies. When the room is empty, we capture a frame to build a initial background model. Then, we update the background model for each region based on the result of head-shoulder detection. Blobs without heads and shoulders inside will be updated into the background model. To simplify the process, we put blobs into

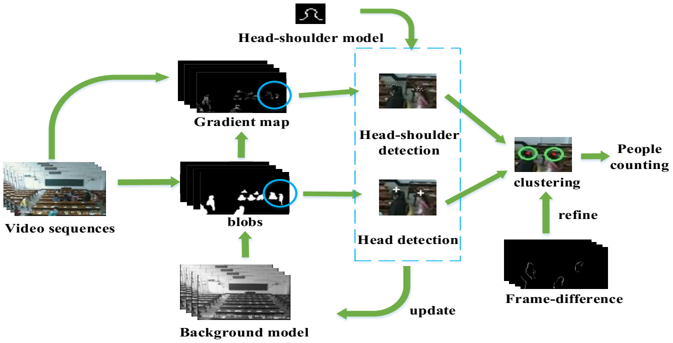


Fig. 1. Overall framework of people counting in indoor environment

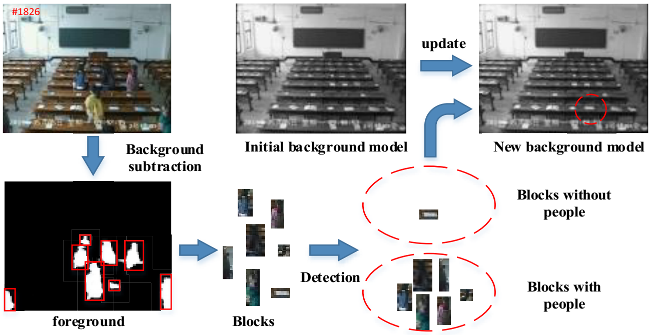


Fig. 2. Block-updating method

some blocks based on the height and width of blobs and blobs are marked by the coordinate of center points of blocks. We update these blocks instead of blobs.

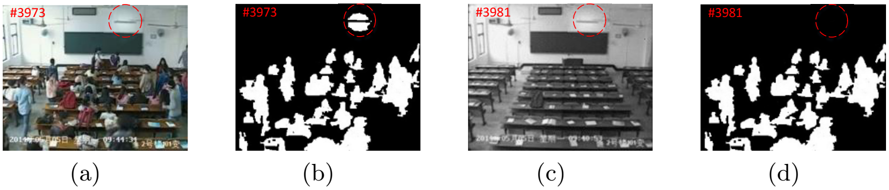
Fig.2 shows the process of block-updating. As shown in Fig.2, because a book was moved in the frame, it became a blob. The foreground was divided into some blocks. The block with the book inside was updated into the background model based on the result of detection. This block-updating way can be described as follows:

$$B_k = B_{k-1} \oplus Pb_j \tag{1}$$

Here B_{k-1} and B_k represent the last background model and the current model respectively. b_j is one block in the foreground. $\oplus$ means update the block b_j to the corresponding area of background model. P is the probability of being updated. We get the probability P by:

$$P = \frac{n}{N} \tag{2}$$

where n is the number of frames one block doesn't contain human bodies. N is the maximum of n and set to 25. Because we always update the no-human pixel into the back ground model, both moving people and stationary people can be extracted. As shown in Fig.3, the blob caused by light is removed after several frames.



(a) (b) (c) (d)

Fig. 3. An example of the block-updating process. (a)The frame when turning on the light; (b)the blob of light bulb is extracted; (c)the area with light bulb is updated to background model after several frames;(d)the blob of light bulb disappeared.

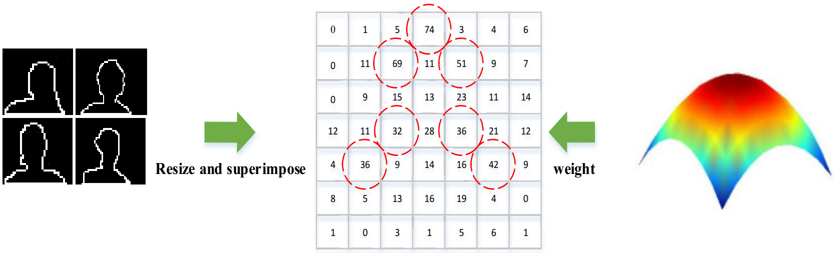
3.2 Head-shoulder Detection

Some blobs we extracted through background subtraction algorithm may contain several people. Zhao and Nevatia [8] proposed a fast and efficient algorithm to locate people in a crowd, focusing on the boundary of the foreground. Similar to [8], head candidates and head-shoulder candidates are detected from foreground images and gradient maps respectively.

Head Candidates. Assuming that heads of humans are visible in a crowd most of the time. Head candidates need meet two conditions: it is the local vertical peak of the boundary, there are enough foreground pixels under it [9]. The result of head detection will be used as candidates for head-shoulder detection.

Head-shoulder Detection. We sample a lot of head-shoulder images from surveillance video and train a set of generic head-shoulder models. Because of the location of cameras, we mainly collect images from two representative categories: back view and side view. Through edge extraction and manually refining, we get a set of profile images as samples.

The training of head-shoulder model is similar to [4]. The training process is a collection of the probabilities of the key points' appearances in the image. But our method process samples more directly and can avoid the error brought by key points selecting. We firstly resize all the samples to a fixed size and superimpose them into a model. The model is divided into some smaller blocks. It's important to note that the top of human head is selected as the reference point. The reference



0	1	5	74	3	4	6
0	11	69	11	51	9	7
0	9	15	13	23	11	14
12	11	32	28	36	21	12
4	36	9	14	16	42	9
8	5	13	16	19	4	0
1	0	3	1	5	6	1

Fig. 4. A toy example of Head-shoulder model

point locates on the upper center of the model, all the head points in samples are aligned to this reference point. As shown in Fig.4, we accumulate the number of edge points in each block. We will get a general distribution of edge points of samples. Then we select some blocks with higher values as the positions of key points. A Parzen Windows method is used to assign every pixel a weight in each block. Assuming the window function obeys a Gaussian distribution, the weight is computed as follows:

$$w(x) = \frac{n}{N}\varphi\left(\frac{x - x_c}{h_N}\right) \tag{3}$$

where n represents the number of edge points in each selected block, N is the number of edge points in samples, x is the coordinate of arbitrary point in a block, x_c is coordinate of the center pixel in the block, h_N is the variance of the Gaussian distribution, φ is the window function, i.e. the Gaussian distribution in our case.

With head-shoulder detection, we can obtain the candidates position of human. We choose K-means clustering with max and min distance to decide the people count. The cluster centers are filtered by checking if there are enough head-shoulder candidates inside. The number of rest cluster centers is treated as the result of counting. Uncertainty of initial center points limits the convergence speed and effect of K-means. Here the sequence of head candidates is utilized as initial center points to make K-means faster and better.

Given a head detect sequence $A = \{\alpha_j\}_{j=1}^{M}$ and a head-shoulder detect sequence $B = \{\beta_j\}_{j=1}^{N}$, $A \cup B$ is the cluster set, and A is initial center points too. The goal is to find cluster center sequence $C = \{\gamma_j\}_{j=1}^{K}$. In order to reduce the impact of occlusions on counting result, we defined a maximum cluster radius R_{max} and a minimum cluster radius R_{min}. R_{min} and R_{max} are set to 1000 and 5000 respectively in our experiments. The algorithm is given in $Algorithm$.1. The normal K-means algorithm will be used to process the rest points in B.

Algorithm 1. Dynamic programming for K-means clustering

```
Input: Head sequences $A = \{\alpha_j\}_{j=1}^{M}$. head-shoulder sequence $B = \{\beta_j\}_{j=1}^{N}$.
Output: cluster center sequence $C$.
for $i = 1; i \leq N; i++$ do
    for $j = 1; j \leq M; j++$ do
        if $distance(\alpha_j, \beta_j) < R_{max}$ then
            delete $\beta_j$;
            $R_{max} = distance(\alpha_j, \beta_j)$;
        else
            $R_{max} = R_{max}$;
        end if
        if $distance(\alpha_j, \beta_j) > R_{min}$ then
            $C \leftarrow \alpha_j$;
        else
            delete $\beta_j$;
        end if
    end for
end for
return $C$;
```

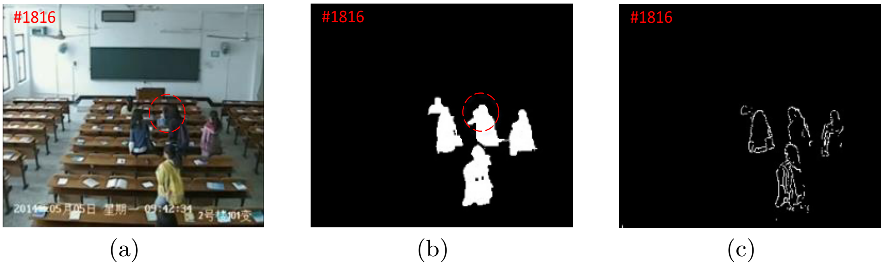
3.3 Temporal Refinement

Except for enter and exit, occlusions is the main cause for change in the count number. We consider reasons of the change as two representative states: One is that the counting number increased as the occlusion disappeared. Another one is the counting number decrease when the occlusion happen. We can find that whatever state in indoor environment, the moving pixels exist. So whether the change is caused by occlusions can be estimated by moving pixels. We mainly consider two cases of occlusions, moving people occluded by stationary people and stationary people occluded by moving people.

As shown in Fig.5, there are two stationary people and four moving people in the frame. A stationary people is occluded by a moving people. we use frame-difference method to get the set of moving pixels. According to different contour of bodies, we divide them into different sequences. When the counting number decreases, we estimate occlusions according to the location of disappeared cluster center. For the case of stationary people occluded by moving people, the position of cluster center will be surrounded by moving pixels in the result of frame-difference. Our judgement is described as follows:

$$I(x_i, y_i) = \begin{cases} 1 \ \text{ if } \begin{cases} min(x_j) \leq x_i \leq max(x_j) \\ min(y_j) \leq y_i \leq max(y_j) \end{cases} \\ 0 \ \text{else} \end{cases} \tag{4}$$

where x_j and y_j is the coordinate of moving pixel in one sequence. x_i and y_i is the coordinate of disappeared cluster center. If $I(x_i, y_i)$ equal to 1, we consider that $I(x_i, y_i)$ is occluded. The counting result remains the same. For the case of moving people occluded by stationary people, occlusions will be estimated by the disappearance of cluster center corresponding moving pixels. Furthermore, if the number of moving pixels is less than a threshold (set to 50 in our experiments) in one frame, we think the state in this indoor environment is steady and keep the counting result the same.



(a) (b) (c)

Fig. 5. Temporal refinement for people counting.(a)occlusion is occurring;(b)some information of head-shoulder disappeared;(c)moving pixels.

4 Experimental Results

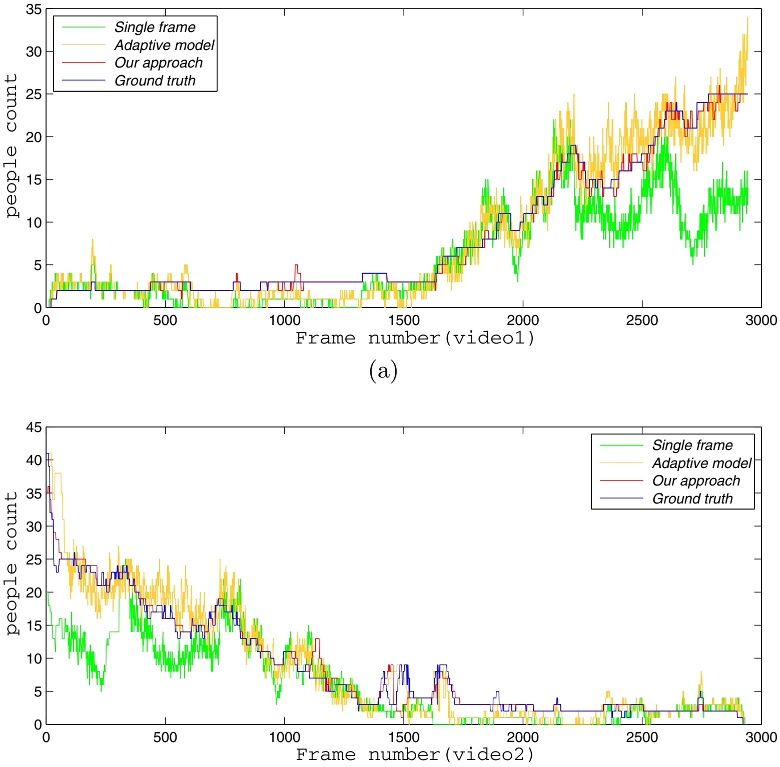
We test our approach on the dataset collected by existing cameras installed on the back wall of classrooms in our university. The dataset reflect all states in these classroom within twenty-four hours. We select videos from two groups respectively, which named *video*1, *video*2.

We compare our approach with [8] and [3] by comparing the counting result on *video*1 and *video*2. The result of [8] is named "single frame", and [3] is "Adaptive model". As shown in Fig.5, because of the limitation of traditional background subtraction in indoor environment, the foreground with some stationary persons inside can't be extracted by [8]. So the counting results of [8] are below the groundtruth during some periods. As shown in Fig.5 (a), in indoor environment, "Adaptive model" [3] is easier to be affected by occlusions and disturbance of noise than our approach. The result of our approach is smoother. Although sometimes the counting result is incorrect, it can get closer to the groundtruth slowly after a number of frames. Because we use spatial temporal information to refine the counting result, the result of our approach falls behind the groundtruth most of the time. If the state in the classroom remains stable for a long time, the refinement on counting result will be constant. Fig.5 (b) shows the same conclusion on *video*2. Fig.6 shows some frames with counting results in our experiments.[1] The evaluation criterion is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(num_g(i) - num_t(i)\right)^2} \tag{5}$$

[1] http://www.nlpr.ia.ac.cn/iva/homepage/jqwang/Demos.htm

where n is the total number of frames in the tested video. $num_g(i)$ is the ground truth of the ith frame. and $num_t(i)$ is the test result of the ith frame. The comparison with [8] and [3] is given in $Table.2$. Comparison in $Table.2$ shows our approach adapts better to the indoor environment.



(a)

(b)

Fig. 6. Counting result on *video*1 and *video*2

Table 1. The comparison results

video name	RMSE		
	Our approach	Single frame [8]	Adaptive model [3]
Classroom1	1.0225	4.9142	2.1847
Classroom2	1.3879	4.9618	2.7618
Classroom3	1.2613	4.9252	2.0137

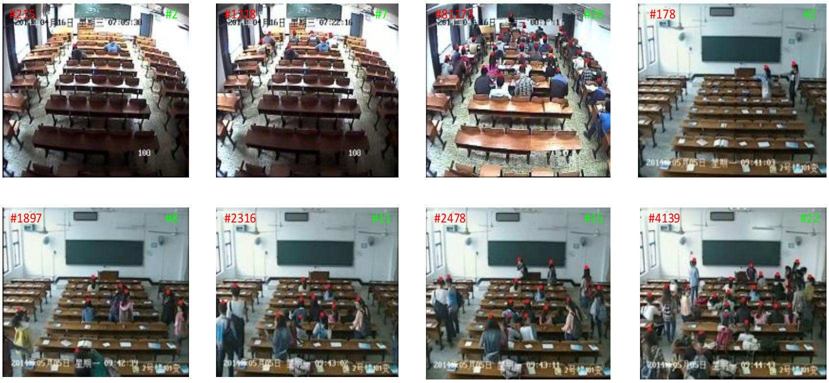


Fig. 7. Counting result in *video*1 and *video*2, the number of human was marked in green on the top right corner of images, and the clustering result was marked on the head of human in red points.

5 Conclusion

In this article, we propose a new method for people counting in indoor environment. We extract foreground objects and present a block-updating way to update the background model, which is fit for the indoor environment. The head-shoulder detection is the key problem in our approach, because it is linked closely with block-updating and clustering. So we trained a generic head-shoulder model and combined with improved K-means clustering to detect human bodies. Based on the spatial temporal information between frames, we refine the counting result and get a stable and accurate number of human in a number of frames. Compared with [8] and [3], our approach achieves a better performance than state-of-the-art approaches in the indoor environment.

Acknowledgements. This work was supported by 863 Program (2014AA015104), and National Natural Science Foundation of China (61273034, and 61332016).

References

1. Chan, A.B., Vasconcelos, N.: Counting people with low-level features and bayesian regression. IEEE Transactions on Image Processing 21(4), 2160–2177 (2012)
2. Li, J., Huang, L., Liu, C.: People counting across multiple cameras for intelligent video surveillance. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 178–183 (2012)
3. Liu, J., Wang, J., Lu, H.: Adaptive model for robust pedestrian counting. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011 Part I. LNCS, vol. 6523, pp. 481–491. Springer, Heidelberg (2011)

4. Sun, C., Zou, Q., Fu, W., Wang, J.: Multiple hypotheses based spatial-temporal association for stable pedestrian counting. In: Huet, B., Ngo, C.-W., Tang, J., Zhou, Z.-H., Hauptmann, A.G., Yan, S. (eds.) PCM 2013. LNCS, vol. 8294, pp. 803–810. Springer, Heidelberg (2013)
5. Teixeira, T., Savvides, A.: Lightweight people counting and localizing in indoor spaces using camera sensor nodes. In: First ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2007, pp. 36–43. IEEE (2007)
6. Wang, J., Fu, W., Liu, J., Lu, H.: Spatio-temporal group context for pedestrian counting. IEEE Transactions on Circuits and Systems for Video Technology, 1–11 (2014)
7. Zhang, E., Chen, F.: A fast and robust people counting method in video surveillance. In: 2007 International Conference on Computational Intelligence and Security, pp. 339–343. IEEE (2007)
8. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II–459. IEEE (2003)
9. Zhao, T., Nevatia, R., Lv, F.: Segmentation and tracking of multiple humans in complex situations. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 2, pp. II–194. IEEE (2001)