Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection

Min Li, Zhaoxiang Zhang, Kaiqi Huang and Tieniu Tan
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
$\{$*mli, zxzhang, kqhuang, tnt*$\}$*@nlpr.ia.ac.cn*

Abstract

This paper proposes a novel method to address the problem of estimating the number of people in surveillance scenes with people gathering and waiting. The proposed method combines a MID (Mosaic Image Difference) based foreground segmentation algorithm and a HOG (Histograms of Oriented Gradients) based head-shoulder detection algorithm to provide an accurate estimation of people counts in the observed area. In our framework, the MID-based foreground segmentation module provides active areas for the head-shoulder detection module to detect heads and count the number of people. Numerous experiments are conducted and convincing results demonstrate the effectiveness of our method.

1. Introduction

Crowd management is a very important task in public places with people gathering and waiting, like bus stations, subway platforms and waiting rooms. Automatic crowd density estimation systems in these scenes can supply useful information for applications such as security surveillance.

Much work has been done on estimation of the number of people in crowded scenes. In [10], the number of people is computed as a function of foreground pixels obtained by background removal using a reference image. But it may fail if background changes. In [4], Haar features [6] are extracted and applied for head detection (shoulder is not included). However, only head is not distinguishing enough for head detection because the appearance and shape of head vary greatly in surveillance scenes. Wu *et al*[12] extract texture features first and then utilize SVM to solve the regression problem of counting people. However, the training process of this method is highly related to specific scenes. Rabaud *et al* [8] propose a method to count crowded moving objects based on clustering a set of extended tracked features. Kilambi *et al* [3] present a heuristic-based and a shape-based method for estimating moving group population. These two methods, however, assume that objects are moving.

A direct method to count people in crowded scenes is to detect the salient omega shapes (head-shoulder shapes). In this paper, a MID (Mosaic Image Difference) based foreground segmentation algorithm is performed first to detect active areas, then a head-shoulder detection algorithm is utilized to detect heads and count the number from the detected foreground areas.

The remainder of the paper is organized as follows. In Section 2, a Mosaic Image Difference (MID) based foreground segmentation algorithm is proposed. Section 3 describes the HOG (Histograms of Oriented Gradients) based head-shoulder detection in detail. Experimental results and analysis are presented in Section 4. Finally, we draw our conclusions in Section 5.

2. MID Based Foreground Segmentation

It is difficult to segment foreground by background modeling methods, like GMM (Gaussian Mixture Model) [9], in places with people gathering and waiting. However, small motions, like people's turning around, wandering and raising heads, surely happen now and then in crowds. It is assumed that these motions approximately satisfy temporal and spatial uniform distributions in a considerably long period of time, because where and when they would happen are completely random. These motions can be effectively represented by the MID (Mosaic Image Difference) feature. Suppose image plane is evenly divided into Mosaic Blocks (MB) with the size of $L_M \times L_M$ (typically $L_M = 4$), the mean RGB vector of MB(m, n) at frame $\#t$ is $M_t(m, n)$, then the MID feature of MB(m, n) at frame $\#t$ can be defined as an indicator function:

$$MID_t(m,n) = \begin{cases} 1 & \text{if } \|M_t - M_{t-1}\|_\infty > T_0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\|\cdot\|_\infty$ denotes the maximum absolute component of a vector; T_0 is a threshold.

By dividing MID series into a number of subsections and normalizing them, the motion occurrence probability of MB(m,n) in each time piece (time of a subsection) of a given period of time can be estimated by:

$$P_l(m,n) = \begin{cases} \frac{\sum_{k \in Sub.\#l} MID_k}{\sum_k MID_k} & \text{if } \sum_k MID_k > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where MID_k represents $MID_k(m,n)$; $l = 1, 2, .., N_S$, and N_S is the number of subsections.

According to the given assumption, if a MB belongs to foreground, its MID series in an observing period of time would satisfy temporal uniform distribution. Three statistics are very important for MID series analysis in a period of time: 1) *MT*: the mean time when motions happened. 2) *VAR*: the variance of time when motions happened. 3) *NNZ*: the number of subsections in which motions happened. *MT* and *VAR* can be computed as:

$$MT(m,n) = \sum l P_l(m,n) \quad (3)$$

$$VAR(m,n) = \sum (l - MT(m,n))^2 P_l(m,n) \quad (4)$$

If the three computed statistics of the MID series of a MB are in the given neighborhoods (parameters related to specific scenes) of their theoretical values, it could be labeled as foreground area.

Those foreground MBs obtained by temporal statistical analysis are just small areas sampled from the crowded areas. According to the given assumption, they would satisfy spatial uniform distribution. So the whole crowded areas can be obtained by the Griding method: 1) evenly divide the whole image into grids with size of $L_G \times L_G$ (typically $L_G = 3L_M$). 2) Label a grid as foreground if there is at least one MB in it being labeled as foreground.

3. HOG Based Head-Shoulder Detection

The most reliable feature for head-shoulder detection in surveillance scenes is its omega-like shape shown in Figure 1 (a). HOG (Histograms of Oriented Gradients) feature has been proven to be a good shape descriptor [1] for human detection. We evaluate several local feature descriptors for head-shoulder detection in the experiment section, and find HOG gives the best results.

Feature Representation Each 32×32 pixel sample is divided into 64 cells with the size of 4×4 pixels, and 4 adjacent cells form a block with the size of 8×8 pixels. There are totally 49 blocks with the overlap of two cells between two adjacent blocks. For each cell in each block, a histogram of 8 orientation bins in 0^o - 360^o is calculated and normalized within this block to represent the local features. The final descriptor is then a vector with the dimension of 1,568. Details about HOG feature extraction can be seen in [1].

4. Experimental Results

4.1 Dataset

Because of lacking of open large dataset for head-shoulder detection, we created one, which is available at [2]. There are 1,755 positive samples with the size of 32×32, together with their left-right reflections (3,510 images in all) for training and 906 (1,812 in all) for testing. Typical head-shoulder samples can be seen in Figure 1 (b). They vary in ethnicities, view angles, appearances and scenes. About two thirds of the positive samples are cropped from two well-known pedestrian datasets, the MIT set [7] and INRIA set [1], and others are cropped from images from the Internet or surveillance videos. For negative samples, most are selected from the INRIA set. For diversity, we add some human-body images without head-shoulders to the negative set. There are 399 head-shoulder-free images with the size of 320×240 for training and 331 for testing respectively.

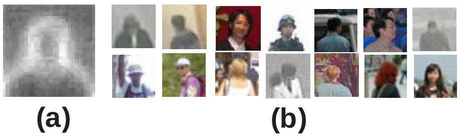


Figure 1. The salient omega feature. (a) Average edge map (like an omega) of head-shoulder samples in the training set (b) Typical head-shoulder samples.

4.2 Training

The positive training set and 11,000 randomly sampled patches from 399 head-shoulder-free images constitute the initial training set, then AdaBoost is used to train the head-shoulder detector. However, collecting a representative set of none-head-shoulder samples is difficult. To overcome the problem of defining this extremely large negative class, a bootstrapping training is adopted. A preliminary classifier is trained on the initial training set, then used to predict the class categories of a large set of patches randomly sampled again from the 399 head-should-free images. False alarms are collected and added to the negative training set for the next iteration of training.

4.3 Testing

Three experiments are conducted. The first two experiments are designed to evaluate the performances of the MID-based foreground segmentation module and the HOG based head-shoulder detection module separately. Then, the combination of the two modules is tested in the third experiment.

In the first experiment, the MID-based foreground segmentation algorithm is tested by a real video (duration:12min) taken from a bus station in the rush hour. Figure 2 (a), (d) and (g) show some selected frames in this video; (b), (e) and (h) show the segmented MBs whose MID series satisfy temporal uniform distribution; and (c), (f) and (i) are the results of the Griding Method. It can be seen that in most cases, our method can exactly detect the crowed areas, no matter how crowded it is.

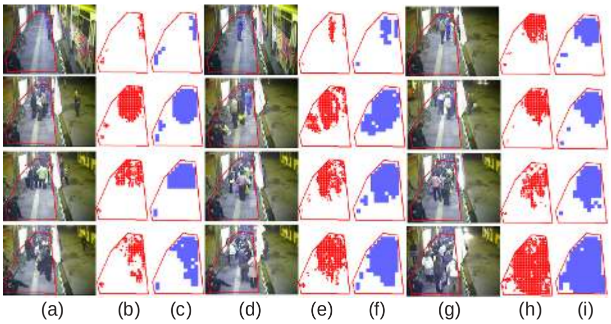


Figure 2. MID based foreground segmentation results in a real scene video.

The second experiment compares performances of the HOG feature and another two popular features : Haar feature [11] and SIFT descriptor [5] for head-shoulder detection. All classifiers are trained on the initial training set (No bootstrapping process) by AdaBoost and evaluated on the testing set (About 331,000 random patches are sampled for the negative set). Results are shown in Figure 3 (a). Apparently, the HOG feature performs much better than the other two features. Dalal [1] mentioned that signed gradients (In 0^o - 360^o) decreases the performance of HOG feature in pedestrian detection. But as shown in Figure 3 (a), signed gradients performs better than unsigned gradients (In 0^o - 180^o) in head-shoulder detection.

Figure 3 (b) shows that the bootstrapping process can decrease the missing rate from about 40% to about 23% at 10^{-4} false positive per window. Figure 4 shows some detection results of our final detector on some surveillance images or daily-life photos.

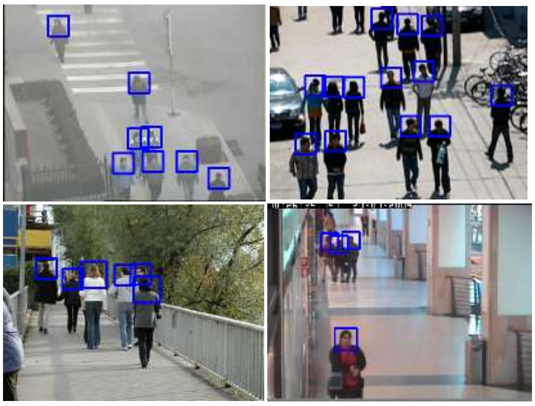
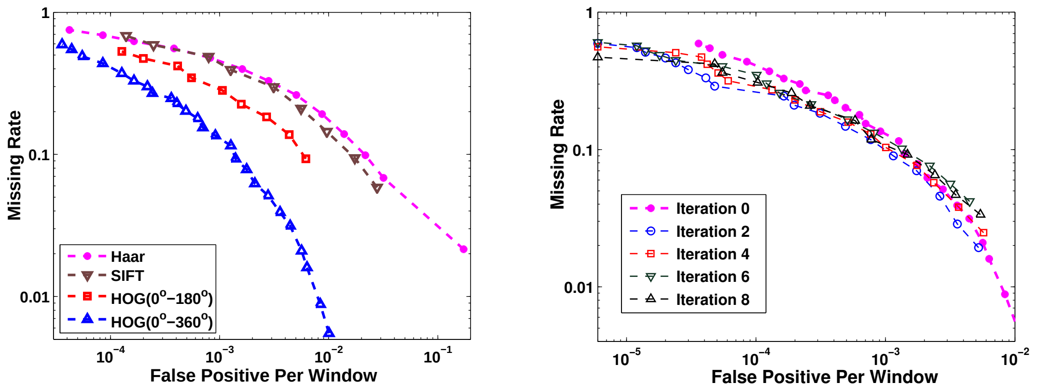


Figure 4. Some detection results of the final head-shoulder detector

In the third experiment, the whole proposed method of estimating the number of people in crowded scenes is tested by a real video taken in another bus station. Figure 5 shows the detection results: (i)-(vi) are segmented foreground maps of some selected frames, blue squares in (a)-(f) are the corresponding head-shoulder shapes detected (The large polygon in red is region of interest). Though the MID-based segmentation is not very accurate, it could decrease the search scope for the head-shoulder detector when there are few people in the observed area. The curve of NOP (number of people) vs. time is shown in Figure 6. As we can see, the number of people detected approximately goes up and down following the ground truth correctly.

5. Conclusions

In this paper, we have proposed a method to estimate the number of people in crowded scenes. This method consists of two modules: a MID based foreground segmentation module to obtain the active areas in the observed area and a head-shoulder detection module to detect the head-shoulder shapes from the detected foreground areas and count the number. This method can not only count the number of people in crowded scenes, but also locate the position of each individual, which has great potential for applications beyond people counting. Experimental results have shown the effectiveness of the proposed method.



(a)

(b)

Figure 3. Performance comparison of (a) different features and (b) different number of iterations in the bootstrapping process

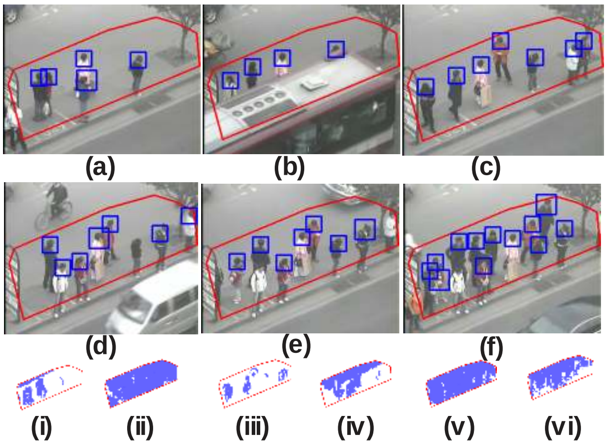


Figure 5. Segmentation and detection results in another testing video.

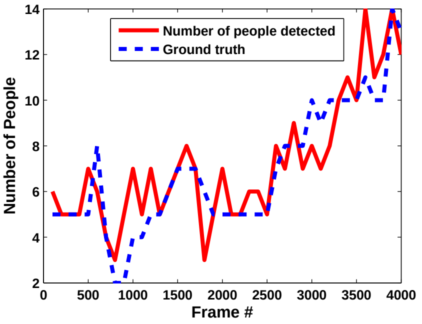


Figure 6. Curve of NOP vs. time for the testing video

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110) and the National Science Foundation (60605014, 60332010, 60335010 and 2004DFA06900). The authors also thank the anonymous reviewers for their valuable comments.

References

[1] N. Dalal and B. Triggls. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] dataset. http://www.cbsr.ia.ac.cn/users/zzx/hsdataset.rar.

[3] P. Kilambi, E. Ribnick, and *etc.* Estimating pedestrian counts in groups. *CVIU*, 2007.

[4] S. Lin, J. Chen, and H. Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans. on Sys., Man, and Cybernetics*, 2001.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[6] M. Oren, C. Papageorgion, and *etc.* Pedestrian detection using wavelet templates. In *CVPR*, 1997.

[7] C. Papageorgious and T. Poggio. A trainable system for object detection. *IJCV*, 38, 2000.

[8] V. Rabaud and S.Belongie. Counting crowded moving objects. In *CVPR*, 2006.

[9] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22, 2000.

[10] S. Velastin, J. Yin, and *etc.* Automated measurement of crowd density and motion using image processing. In *Proc. of IEEE Conf. on Road Traffic Monitoring and Control*, 1994.

[11] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[12] X. Wu, G. Liang, and *etc.* Crowd density estimation using texture analysis and learning. In *Proc. of IEEE Conf. on Robotics and Biometrics*, 2006.