HUMAN DETECTION IN CROWDED SCENES

Ya-Li Hou and Grantham K. H. Pang

Industrial Automation Research Laboratory,
Department of Electrical and Electronic Engineering, The University of Hong Kong

ABSTRACT

In this paper, our focus is to segment the foreground area for human detection. It is assumed that the foreground region has been detected. Accurate foreground contours are not required. The developed approach adopts a modified ISM (Implicit Shape Model) to collect some typical local patches of human being and their location information. Individuals are detected by grouping some local patches in the foreground area. The method can get good results in crowded scenes. Some examples based on CAVIAR dataset have been shown.

A main contribution of the paper is that ISM model and joint occlusion analysis are combined for individual segmentation. There are mainly two advantages: First, with more sufficient information inside the foreground region, even the individuals inside a dense area can also be handled. Secondly, the method does not require an accurate foreground contour. A rough foreground area can be easily obtained in most situations.

Index Terms— Human detection, Occlusions, Implicit Shape Model

1. INTRODUCTION

People counting and human detection are two important problems in visual surveillance. It is useful for shopping mall managers to get the knowledge about the number of people in the mall each day. For the safety of people and facilities, video surveillance has become more and more important. Detecting individuals is usually the first step for further video analysis.

Some substantial work on human detection has been carried out. However, situations with significant occlusions are still an open problem.

2. RELATED WORK

People counting and human detection has become a hot topic in these years. All the methods may be classified into two categories. The first one assumes that a foreground area for the crowd has been obtained. People counting and detection are achieved by segmenting the foreground into individuals, like [1-4]. The other category exhaustively searches an image with a scanning window. Each window is classified as human or non-human based on shape, color or motion features [5-8]. To reduce the number of scanning windows, Li et al. [9] search the head-shoulder shape based on the HOG (Histogram of Oriented gradients) only inside the foreground region. Most methods in this category are computationally expensive and only work for a crowd with slight occlusions. Our method belongs to the first category and only the most related work in the first category will be discussed in this section.

Zhao and Nevatia [1] segment foreground area using head detection. In their early work, head is detected by checking local peaks on foreground contour. In their later work [2], a simple 'Ω' template is also considered for head detection inside the foreground area.

Rittscher et al. [3] sample some informative feature points from foreground contour and label them as top, bottom, left and right based on their local contour information. A variant of EM (Expectation-Maximization) algorithm is used to find the best grouping of the points with rectangles. In the E-step, the assignment of points to rectangles is obtained based on their distance to the corresponding top, bottom, left or right borders. In the M-step, rectangle sizes, locations are adjusted based on the association property of points in the E-step. Points with low assignment probability have low influence on the rectangle.

Although [3] has tried to reduce the requirements for foreground contour by only sampling some feature points from the contour, both [1, 2] and [3] rely on a good foreground contour, which cannot be easily satisfied in most situations. Especially when the crowd shows only occasional slight movement, it is very difficult to get a very good foreground contour. Besides, with little information inside the foreground area, both [1, 2] and [3] can hardly handle the high ambiguity at the center of dense crowds.

In [4], clustering corner-like points in foreground area is attempted to handle a challenging situation. This method does not rely on informative foreground contours. Some good individual detection results have been achieved in a large crowd. However, with only the locations of corner-like points, it is hard to cluster the points well in a dense area.

On the other hand, ISM (Implicit Shape Model) used in [10] collects a codebook associated with the local patches and their locations with respect to human center. The codebook contains more sufficient information than only corner-like points. Each patch votes for the human centers and Mean shift algorithm is used to find the maxima in the 3D voting space. However, a person with significant

occlusions may not be able to get enough votes. Only results with slight occlusions are shown in the paper.

This paper assumes that the foreground region has been detected and aims to develop an approach to segment the foreground region to individuals. A codebook is established to collect human local patches and their location information. Individual segmentation is formulated as a problem to group local patches with some rectangles. The details of the method will be introduced in Section-3.

3. THE METHOD

The approach includes two stages. In the training stage, a codebook consisting of some typical human local patches and their location information is collected. In the testing stage, the codebook will provide location information for the extracted local patches in a test image. Individual segmentation is performed based on the patches.

3.1. Training stage

The training images should contain some fully-visible human beings. It is assumed that we have the foreground region of the selected persons for training. Also, rectangles have been placed on those persons to indicate their locations.

Step-1: Patch extraction. First, a scale-invariant DoG (Difference of Gaussian) interest point detector [11] is performed on all the training images. This detector gives the location and scales of each detected interest point. By using the foreground region mask, only the interest points from the selected persons will be used. Next, local patches with a radius of three times the detected scales are extracted around each point. All the patches are resized to 25*25 in our evaluations. A number of patches can be collected based on all the training images.

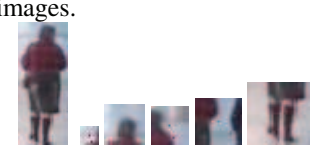

Fig.1. An example of a person and some extracted local patches.

Step-2: Patch clustering. To limit the codebook size, all the extracted patches in step-1 will be clustered into categories. Agglomerative clustering in [12] has been used in our evaluations. Starting with each patch as a separate cluster, two clusters with the smallest similarity distance are combined in each iteration. NGC (Normalized Grayscale Correlation) in equation (1) is used as the similarity measure of two patches with 25*25 pixels. The similarity measure of two clusters, C_1 , C_2 are obtained in (2). The process continues until the smallest similarity distance is below 0.4. In (1) and (2), q_l , $l = 1, 2$ are two local patches, $\overline{q}_l$ is the average value of the patch, $q_l^{x,y}$ is the pixel at the xth row, yth column in the patch. $|C_1|$ and $|C_2|$ are the number of patches in cluster C_1 , C_2 respectively.

$$NGC(q_1, q_2) = \frac{\sum_{x,y=1,2,...25} (q_1^{x,y} - \overline{q}_1)(q_2^{x,y} - \overline{q}_2)}{\sqrt{\sum_{x,y=1,2,...25} (q_1^{x,y} - \overline{q}_1)^2 \sum_{x,y=1,2,...25} (q_2^{x,y} - \overline{q}_2)^2}} \quad (1)$$

$$similarity(C_1, C_2) = \frac{\sum_{q_1 \in C_1, q_2 \in C_2} NGC(q_1, q_2)}{|C_1| * |C_2|} \quad (2)$$

Step-3: Codebook formation. The cluster centers in step-2 are saved as the codebook entries. Next, we need to collect location information for each entry.

Each patch from step-1 has to register its location with all the matched entries (where NGC between the local patch and the code is above 0.4). As mentioned before, each selected person has been annotated with a rectangle in all our training images. The patch locations are registered based on a 3*3 block as shown in Fig. 2. Hence, nine (3*3) spatial occurrence values can be collected for each codebook entry. Finally, the collected location information for each entry is normalized such that $\sum_i p_i = 1$, where p_i , $i = 1, 2, ...9$ is the probability of the entry occurs in block i.

Fig.2. A person is divided into 3*3 blocks, which is used to indicate the patch location.

After the training stage, a codebook consisting of N entries and their spatial occurrence probabilities in each block is established, as shown below.

Codebook entries	*Spatial occurrence probability in each block*				
c_1	p_{11}	p_{12}	p_{13}	…	p_{19}
c_2					
…					
c_N	p_{N1}	p_{N2}	p_{N3}	…	p_{N9}

3.2. Testing stage

Similar to section 3.1, the scale-invariance DoG interest point detectors are performed and multi-scale local patches are extracted from a test image. It is supposed that background has been removed. Only the interest points inside the foreground area will be used.

Step-1: Information collection. This step would collect location information for all the patches in the test image.

For each patch, all the codebook entries are searched. The matched entries (NGC between the codebook entry and the patch is above 0.4) will cast a weighted vote for the location of the patch based on their similarity. A straight forward way is to use NGC between the extracted patch and the codebook entry as the vote weights. Then, the

probability of patch q_l in the ith block, p_{li}, can be obtained with equation (3). p_{ni} is the probability of the code entry, c_n, in the ith block, which has been saved in the codebook. In this way, a 3*3 location table can be established for each extracted local patch. Fig. 3 has shown an example of two location tables. It can be seen that the point around the head area has a high probability in block 4 while the one around the feet area has a high probability in block 2 and 3.

$$p_{li} = \frac{\sum_{NGC(q_l,c_n)>0.4} NGC(q_l,c_n) * p_{ni}}{\sum_{NGC(q_l,c_n)>0.4} NGC(q_l,c_n)}, \quad i = 1,2,...9 \tag{3}$$

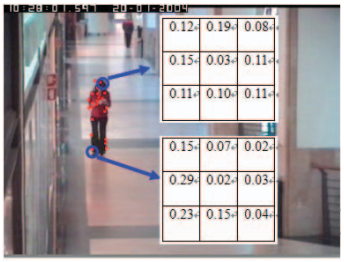

Fig.3. A location table is collected for each extracted local patch.

Step-2: Individual detection. First, a set of initial human candidates are nominated. In our evaluations, a simple rectangle is used as the human model. For each detected interest point with a sufficient location probability in block 4, a candidate rectangle is nominated with that point as the middle of the upper border. Average human size is used and it varies as a function of y-coordinates due to perspective distortion. The candidates with small overlap with foreground are removed. Initial nominated rectangles are $R = \{r_k, k = 1...K\}$, K is the number of rectangles.

Given a specific configuration, a binary matrix, M, is used to indicate the assignment of each local patch to the candidate rectangles. $M = \{m_{lk}\}$, $l = 1...L$, $k = 1...K$. L is the total number of local patches. If the interest point l is within the un-occluded region of rectangle k, then $m_{lk} = 1$, otherwise, $m_{lk} = 0$. Usually, it can be assumed that human candidates with smaller y-coordinates will be occluded by those with larger y-coordinates.

Based on the assignment, a score can be used to evaluate the crowd configuration. In our evaluations, the score for $R = \{r_k, k = 1...K\}$ is defined as

$$s = \sum_{l=1:L} \sum_{k=1:K} m_{lk} p_{li}^k . \tag{4}$$

p_{li}^k is the probability of point l in block i of rectangle k,. i can be 1, 2, … 9, and it depends on the location of the point in the rectangle.

Starting from the initial set of candidate rectangles, the best configuration with the largest score is obtained by repeatedly removing the candidates one by one. The details of the implementation of the algorithm are listed as follows. Close candidates are merged in the final results.

Algorithm for Individual Detection

Initialization:

Initial rectangles are nominated. All the rectangles are sorted in descending order according to their y-coordinates. $R = \{r_k, k = 1...K\}$

The initial score, s_0, is obtained by (4).

Loop until the rectangles are not changed.

Iterate k=1…K

(a) $R' = R - r_k$;

(b) Assign each point to the rectangle in R'. Each point can only be assigned to one rectangle.

(c) Calculate the score s with (4).

(d) If s>= s_0, then $R \leftarrow R'$ and $s_0 = s$.

(e) k=k+1;

K= the number of remaining rectangles in R.

Output:

The number of rectangles, K;

The location of each rectangle, $\{x_k, y_k\}$, $k = 1,2,...K$.

4. EVALUATIONS

The CAVIAR dataset [13] is a commonly used video set for human detection. It was taken by a stationary camera fixed at a few meters above the ground. The image size is 384*288. The ground truth of each frame has been provided with human individuals annotated in a rectangle. All the video sequences we have used are from the shopping center sequences with the corridor view.

Training set: The training images were extracted from five video sequences in the CAVIAR dataset. Twenty-two images with 10 persons were used in our evaluations to form the codebook. Only fully-visible persons with a certain size were used. The rough foreground region for the selected training images was obtained manually. The blue rectangles in Fig. 4 were from the annotations of the CAVIAR dataset.

Fig.4. Examples of training images.

Testing set: The test images are from two video sequences different from the training set in the CAVIAR dataset.

In our evaluations, SIFT key point detector Version-4 [14] was used as the interest point detector. The other programs were implemented in MATLAB. Even with a small training set, some good results have been achieved. Fig. 5 shows some examples of the results.

Fig.5. Examples of results on the CAVIAR dataset. Upper row: initial rectangles; bottom row: final results.

It can be observed that most individuals have proposed more than one rectangle candidates during initialization. Although only a simple rectangle model was used in our evaluations, most human beings have been detected well. The method has also shown a good performance for the significantly occluded individuals.

With a more accurate human model, the assignments of patches to each individual will be more accurate. The individual locations can be localized more accurately. In addition, more precise location information (with more blocks in section 3.1, step-3) can also improve the localization results.

5. CONCLUSIONS

In this paper, a method to segment foreground area into individuals is presented. In the future, more tests will be performed on other video sequences. To handle more complicated scenes, a larger training set is necessary.

For future work, lots of improvements can be considered. Currently, human with the same y-coordinate is set as the same size based on perspective distortion. Later, human size for each person will also be adjusted in the individual detection step.

In the current method, the use of appropriate local patch descriptors and similarity measure are important. Alternate local patch descriptor like HOG(Histogram of Oriented Gradients) [5] should also be examined to improve the detection performance.

[1] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *PAMI,* vol. 26, pp. 1208-1221, 2004.

[2] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and Tracking of Multiple Humans in Crowded Environments," *PAMI,* vol. 30, pp. 1198-1211, 2008.

[3] J. Rittscher, P. H. Tu, and N. Krahnstoever, "Simultaneous estimation of segmentation and shape," in *CVPR*, 2005, pp. 486-493.

[4] Y.-L. Hou and G. K. H. Pang, "Human Detection in a Challenging Situation," in *ICIP*, 2009.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886-893.

[6] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," in *CVPR*, 2006, pp. 1491-1498.

[7] O. Tuzel, F. Porikli, and P. Meer, "Human Detection via Classification on Riemannian Manifolds," in *CVPR*, 2007, pp. 1-8.

[8] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling," in *ICCV*, 2009.

[9] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *ICPR*, 2008, pp. 1-4.

[10] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR,* 2005, pp. 878-885.

[11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV,* vol. 60, pp. 91-110, 2004.

[12] B.Leibe, A.Leonardis, and B.Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV'04 Workshop on Stat. Learn. in Comp. Vis.,*, 2004, pp. 17-32.

[13] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[14] http://people.cs.ubc.ca/~lowe/keypoints/.