CROWD COUNTING AND SEGMENTATION IN VISUAL SURVEILLANCE

Lu Wang, Nelson H.C. Yung

Department of Electrical and Electronic engineering, The University of Hong Kong

ABSTRACT

In this paper, the crowd counting and segmentation problem is formulated as a maximum a posterior problem, in which 3D human shape models are designed and matched with image evidence provided by foreground/background separation and probability of boundary. The solution is obtained by considering only the human candidates that are possible to be un-occluded in each iteration, and then applying on them a validation and rejection strategy based on minimum description length. The merit of the proposed optimization procedure is that its computational cost is much smaller than that of the global optimization methods while its performance is comparable to them. The approach is shown to be robust with respect to severe partial occlusions.

Index Terms— Crowd counting, crowd segmentation, model based segmentation, Bayesian method

1. INTRODUCTION

Crowd counting and segmentation is an important, yet challenging problem in video surveillance. The main difficulty resides in the partial occlusion that prevalently exists in the crowd.

Previous work to human crowd segmentation can be classified into two categories: motion based approaches and shape based approaches. The argument of the motion based approaches [1, 2] is that the motion field of the same human object is relatively uniform. Hence low level features are extracted and tracked, and the features with similar trajectories are clustered to form human objects. However, without explicit shape constraint, the performance of this kind of methods degrades when multiple human objects have similar trajectories.

Shape based methods usually detect and segment crowds in a single image. In [3], edgelet features are introduced to construct boosted body part classifiers and responses of part detectors are combined to form a joint likelihood model of human. In [4], local information from image patches is collected in a probabilistic Hough voting procedure. However, these two methods become problematic under crowded and cluttered environment. Instead of learning the human shape, 2D [5] or 3D [6] models can be explicitly designed to represent human shapes. In [5], a hierarchical part-template matching is proposed to handle partial occlusions. However, as the optimization algorithm is greedy, the result may be wrong because the assumed occlusion order may not be correct. In [6], Markov Chain Monte Carlo (MCMC) is used to efficiently search the solution space. However, this kind of random search can be computationally expensive. Expectation Maximization (EM) is used in [7, 8] to assign image features to human candidates, and occlusion reasoning is performed in the M-step .

This paper proposed a Bayes approach that, given the foreground and camera parameters, segments the crowd into individuals. Considering that finding the global optimal solution that maximizes the posterior probability is very computationally expensive [6-8], while the assumed occlusion order might be wrong of the faster greedy approaches [3, 5], in our implementation, the optimization is performed within a portion of the candidates in each iteration. The merit of the proposed optimization procedure is that its computational cost is much smaller than the global optimization methods while its performance is comparable to them.

2. PROBLEM FORMULATION

Our goal is to segment the foreground into individual human objects where occlusion may exist. We formulate the crowd counting and segmentation problem as a maximum a posterior (MAP) problem such that the optimal solution θ^* is given by

$$(\theta^*) = \arg\max_{\theta} P(\theta \mid I) \tag{1}$$

where θ consists of the number of human objects n and their corresponding models (m_i, i=1,…,n); I is the foreground mask. According to Bayes Rule, (1) can be decomposed into a prior term and a joint likelihood term.

$$P(\theta \mid I) = P(\theta)P(I \mid \theta) / P(I) \propto P(\theta)P(I \mid \theta) \tag{2}$$

We assume that the prior of a solution is the product of the prior probabilities of each individual human object and is defined as

$$P(\theta) = \prod_{i=1}^{n} P(\mathbf{L}_i)P(\mathbf{L}_i \mid \mathbf{L}_{-i}) \tag{3}$$

The first term of (3) gives each human object in θ a penalization according to their real world position $\mathbf{L}_i$, avoiding n to be extremely large, and it is defined as

$$P(\mathbf{L}_i) = \exp(-\alpha(\mathbf{L}_i)) \tag{4}$$

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

where α is a quadratic function of the distance between $\mathbf{l}_i$ and the camera – the larger the distance is, the smaller $\alpha(\mathbf{L}_i)$ is, which considers the perspective view effect of the imaging. The second term is the prior probability of the i^{th} human object's position relative to others (denoted as $-i$). It represents our prior knowledge that two persons must keep a certain distance away from each other in the real world and is given by

$$P(\mathbf{L}_i \mid \mathbf{L}_{-i}) = f(\min_{j\in 1,\ldots,n, j\neq i} \left|\mathbf{L}_i - \mathbf{L}_j\right|)$$
$$f(d) = \begin{cases} d/d_{min} & \text{if } d \leq d_{min} \\ 1 & \text{if } d > d_{min} \end{cases} \tag{5}$$

where d_{min} is the minimum distance required for any two human objects. Assuming the pixels are independent, the likelihood is defined as

$$P(I \mid \theta) = \prod\nolimits_{k\in I} P(k \mid \theta) = \exp(-\sum_{k\in I}(1 - S_\theta(k)) \tag{6}$$

where $S_\theta(k)$ is the score of matching the un-occluded part of the boundary of m_i with the foreground probability of boundary (*pb*) [9], if k belongs to the un-occluded part of m_i; otherwise $S_\theta(k) = 0$, meaning that pixel k is not inside any human object models of θ.

3. IMPLEMENTATION

We first extract the foreground from the input image by a multiple adaptive thresholds method [10] and obtain the camera parameters by [11]. Then an upper semi circle detector is used to give an exhaustive nomination of candidates. To find the global optimization solution is computationally expensive. Therefore, in our implementation, by analyzing the mask and the relationship between candidates, in each iteration, only a group of the possible un-occluded candidates are selected for model matching, and the results are fed into a minimum description length (MDL) based validation and rejection procedure. This kind of optimization reduces the computational cost significantly without sacrificing much of the performance.

3.1 Candidate nomination

From our observation, the only reliable feature of a human is the head. Therefore, we use a head detector to provide the candidate nomination. The applied method [12] is a Hough-like circle detector, in which each boundary element spreads its vote, modulated by the edge magnitude, into (x_c, y_c, r), the postulated circle's center and radius. Because the boundary elements contain orientation information, they vote for circle centers only if they are tangential to the circle. In our application, to detect upper semi circles, horizontal or slanted edge response just votes for the circle center below it, while vertical response of edge would vote for both its left and right circle centers. The directional filter we use is *pb*, which effectively removes the edge response of textures and thus makes the number of false positive detection of heads much smaller. The scale set of the circle detection *Rad* is determined by projecting two spheres, representing the lower and upper bounds of human head size respectively, onto the input image and taking half of the projections' widths as the bound r_{min} and r_{max}.

Having the upper semi circle detection response of each radius in *Rad*, the maximum of the responses of different radii for each pixel (x, y) of I is obtained to get the final response $R(x, y)$. Then the local maxima of $R(x, y)$ that are above a threshold are found to obtain the head candidates set C. Redundant candidates are removed: 1) if the center of one circle is inside another circle, the one with the weaker response is discarded; 2) those candidates that do not have enough foreground area below them are also discarded. An example of head candidate detection is shown in Fig. 1.

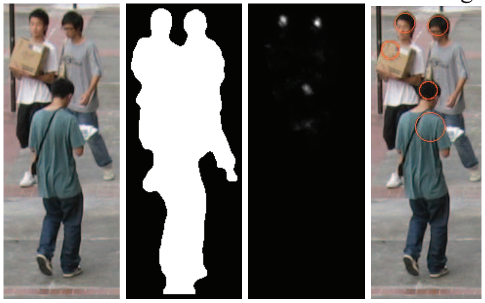

Fig.1 Head candidate detection: (a) The input image; (b) The foreground mask; (c) The head detection response $R(x, y)$; (d) Detected circles overlaid on the input image.

3.2 Candidate selection for model matching

Ideally, only the currently un-occluded candidates should be selected for model matching in each iteration (by currently un-occluded candidates, we mean that the candidates are either un-occluded or if they are occluded, their occluding candidates have been validated and removed from the mask in the last iteration). This will guarantee that the model matching can be executed correctly.

In the first iteration, we aim to find the possibly un-occluded human objects. Assume that the full body is within the mask for all the human objects. Lower extrema (LE) of the mask boundary are extracted. Then, restricting LEs' vertical and horizontal distances with the head top, LEs that can represent a candidate's feet are assigned to the candidate. Only the candidate that is assigned at least one LE is considered to be possibly un-occluded. For the following iterations, bounding boxes are drawn for each candidate. A bounding box is a quadrangle that defines the maximum possible range of a human being, given the head top position. If a candidate's bounding box's intersection with the current mask $I_{cur}= I - I_{occ}$ (I_{occ} is the occupancy map generated by the validated candidates) is different from its intersection with the mask in the last iteration, meaning that the surrounding candidates' status has been changed, it is selected for model matching in the current iteration.

3.3 Model matching

The 3D human shape model we use consists of seven parts – the head (modeled by an ellipsoid), the shoulder (a half ellipsoid), the torso, the left/right thigh and the left/right calf (each by a cylinder) – as is shown in Fig. 2. To restrict the searching space, only 17 walking/standing postures are selected for model matching. In addition, seven orientations (0, 30°, 60°, 90°, 120°, 150° and 180°) and four scales (corresponding to the height of 1.55m, 1.65m, 1.75m and 1.85m respectively) are used. Further, the head is allowed to deviate horizontally from the vertical torso center by $\pm r_{head}$ /2 and $\pm r_{head}$ /4 where r_{head} is the radius of the model head.

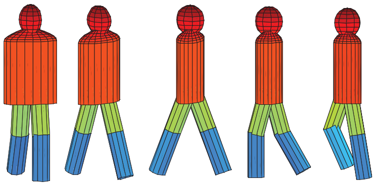

Fig. 2 Illustration of some 3D human models

Given a selected candidate c_i, each model M_j in the model set is projected onto the image according to the candidate's head top position and the camera parameters. The matching is measured by both the model's region coverage with the current mask I_{cur} and the model boundary's matching with the *pb* map of the foreground. The *pb* map used here is not treated with the non-maximum suppression; therefore, it is similar to the distance transform performed on an edge map. The matching score $S(M_j)$ is given by the product of the region matching score $S_r(M_j)$ and the shape matching score $S_s(M_j)$

$$\begin{aligned} &S(M_j) = S_r(M_j)S_s(M_j) \\ &S_r(M_j) = \Gamma(M_j \,\&\, I_{cur}) - \Gamma(M_j \,\&\, (1-I)) \\ &S_s(M_j) = \sum\nolimits_{k \in Mb_j \& I_{cur}} pb(k) / \Gamma(Mb_j \,\&\, I_{cur}) \end{aligned} \tag{7}$$

where $\Gamma(\cdot)$ denotes the size of the non-zero pixels of the image and Mb_j is the boundary image of M_j. In $S_r(M_j)$, the minuend encourages larger area to be explained by the model while the subtrahend penalizes the regions falling out of I and hence prevents extremely large models to be selected. $S_s(M_j)$ is simply the average *pb* value of the un-occluded part of the model boundary. The best matched model m_i is then selected as the one that results in the maximum increase of the posterior as followed:

$$m_i = \arg\max_{M_j}(S(M_j) + \log P(\mathbf{L}_j \mid \mathbf{L}_{val})) \tag{8}$$

where $\mathbf{L}_{val}$ represents the real world positions of the validated candidates and $P(\mathbf{L}_j \mid \mathbf{L}_{val})$ is defined in (5). As the penalty term $P(\mathbf{L}_i)$ is nearly a constant for the same head candidate, it is not included in (8).

3.4 Candidate validation and rejection

The candidate that has good model matching quality (high matching score), indicating that the candidate is unlikely to be a spurious candidate, and the candidate that is the nearest to the camera, indicating that the candidate is unlikely to be un-occluded, are preferred to be validated. To combine our preference for these two kinds of candidates, we propose the following candidate validation and rejection procedure. The idea is that we first reject the candidates that have unsatisfactory model matching quality and the candidates whose corresponding area can be better explained by other candidates, and then confirm the candidate that is not occluded by any other candidates.

a) Consider single candidate's model matching quality

For each candidate c_i that is selected in 3.2, if adding m_i into θ cannot increase the posterior $P(\theta \mid I)$, c_i is rejected. This situation occurs when m_i is poorly matched or m_i just explains a relatively small area.

b) Consider other candidates' model matching quality

For each remaining candidate c_i and the corresponding model m_i, the MDL principle is applied to evaluate if it should be rejected or not. The evaluation is in terms of the *savings* that can be obtained by rejecting c_i:

$$\begin{aligned} &Sav_i = SE_i - SE_{-i} + SM_i \\ &SE_i = \Gamma(m_{cur,i})(1 - S_s(m_i)) \\ &SE_{-i} = \sum\nolimits_{k \in m_{cur,i}} (1 - \max_{j \neq i}(S_s(m_j, k))) \\ &SM_i = \alpha(\mathbf{L}_i) \end{aligned} \tag{9}$$

where $m_{cur,i}$ is m_i's intersection with I_{cur}. SE_i is the error introduced by using m_i to explain $m_{cur,i}$. SE_{-i} is the error introduced by combining other candidate models matched in the current iteration to explain $m_{cur,i}$. $S_s(m_j, k) = S_s(m_j)$ if $k \in m_{cur,j}$ and $S_s(m_j, k) = 0$ otherwise. SM_i is the cost of the model and α and $\mathbf{L}_i$ are the same as they are defined in (4). If Sav_i is positive, c_i is rejected.

After rejecting the candidates that are not good enough, the candidates, which are nearer to the camera than any other candidates that intersect with them, are validated. The validated candidates are then added to θ, and their explained regions are subtracted from I_{cur} and added to I_{occ}. The entire optimization procedure is summarized below.

```
**Algorithm:** optimization algorithm
Given the candidate nomination $C$,
**initialize** $\theta = \emptyset$, $I_{occ}$ as empty (black image), the validated candidates set $C_{val} = \emptyset$, the rejected candidates set $C_{rej} = \emptyset$, and the posterior as $P(\theta \mid I) = \exp(-A(I))$.
**while** $C_{val} \cup C_{rej} \neq C$
    1. Select the possible currently un-occluded candidates. (3.2)
    2. For each candidate selected in step 1, perform model matching and select the best matched model as the one increases the posterior most. (3.3)
    3. Reject and validate these candidates and update $C_{rej}$, $C_{val}$, $\theta$, $I_{cur}$, and $I_{occ}$. (3.4)
**end**
**return** $\theta$.
```

4. EXPERIMENTAL RESULT

The proposed method is evaluated on 70 images selected from 23 image sequences that are taken around our campus. These images contain crowded scenarios where severe partial occlusion of human exists and there are totally 486 human objects in them. The parameters of the proposed method are fixed for all the test image sequences. To nominate nearly all the true candidates, the threshold for head detection is set a bit lower to be 0.1. To determine $\alpha(\mathbf{L}_i)$, we require that a human object have at least the area of a head visible on the image and this area should match satisfactorily. Therefore, $\alpha(\mathbf{L}_i)$ is set to be the average image head area of a human located at $\mathbf{L}_i$ times 0.5 (the required matching score). The minimum distance d_{min} between two human objects is set to be $0.4m$, according to the analysis of pedestrians' inter-distance in the real world.

Several segmentation results are shown in Fig.3 with the best matched models' boundaries overlaid on the human objects. Due to the space limit, not the whole image but only the portion that contains the occlusion is shown here. It can be seen that, from Fig. 3(a) to Fig. 3(f), crowd counting is correctly performed, while in Fig. 3(g), the girl carrying a brown bag is not detected. Wrong posture and orientation estimation also exist [e.g., the occluded girl in Fig. 3(a) and the girl in black in Fig. 3(f)], and in Fig. 3(g), both the position and posture of the girl in white are wrongly estimated, caused by the incorrect head detection.

Totally, among the 486 human objects in the 70 images, 465 (95.6%) are detected and 17 (3.5%) false alarms are produced by the proposed method, which is comparable with the global optimization results shown in [6] and [7]. Because we use both the region and edge information, whereas [6] just uses region and [7] just uses bounding contour of the silhouette, our method is expected to have a higher accuracy than [6] and [7] in model matching.

5. CONCLUSION

A Bayesian approach for crowd counting and segmentation has been proposed in this paper. Foreground and probability of boundary are used to provide image evidence. The solution is obtained in a way that balances the computational cost and the performance. Results on challenging data show the robustness of the proposed method.

However, there are still missed detections, false alarms and wrong posture estimation. To improve the performance, the most important future work is to combine the crowd segmentation results across consecutive frames, which can resolve the ambiguities of a single frame, to obtain a more reliable counting, segmentation and posture estimation performance.

6. ACKNOWLEDGEMENT

This work was supported in part by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China, under Project HKU719608E and in part by the Postgraduate Studentship of the University of Hong Kong.

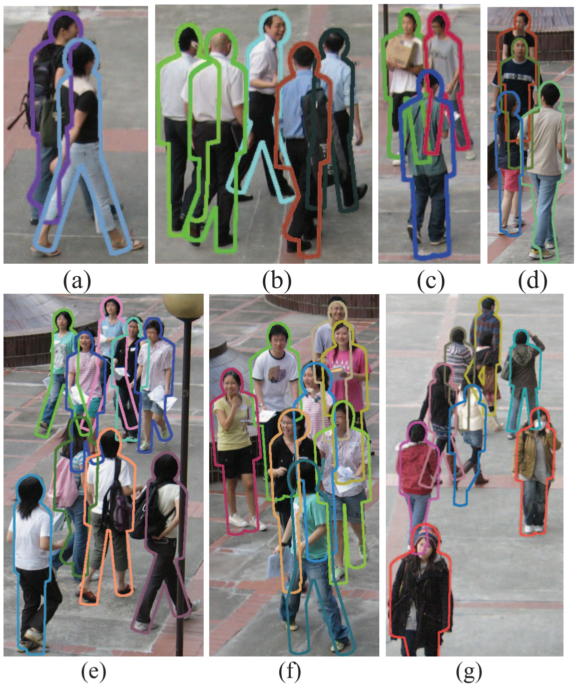

Fig. 3 Experimental result of crowd counting and segmentation

7. REFERENCES

[1] G. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," *CVPR*, 2006, pp. 594-601.

[2] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *CVPR* 2006, pp. 705-711.

[3] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *CVPR* 2005, pp. 90-97.

[4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR* 2007, pp. 878 - 885.

[5] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *ICCV* 2007, pp. 1-8.

[6] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *CVPR* 2003, pp. 459-266.

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

[8] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentaton," *ECCV* 2008, pp 691-704.

[9] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE TPAMI,* vol. 26, no. 5, pp. 530-549, 2004.

[10] L. Wang and N. H. C. Yung, "Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation," *IEEE TITS*, vol. 10, no. 4, 2009.

[11] X. C. He and N. H. C. Yung, "New method for overcoming ill-conditioning in vanishing-point-based camera calibration," *Optical Engineering,* vol. 46, no. 3, 037202, 2007.

[12] S. Bileschi and L. Wolf, "Image representation beyond histograms of gradients: The role of Gestalt descriptors," *CVPR* 2007, pp. 1-8.