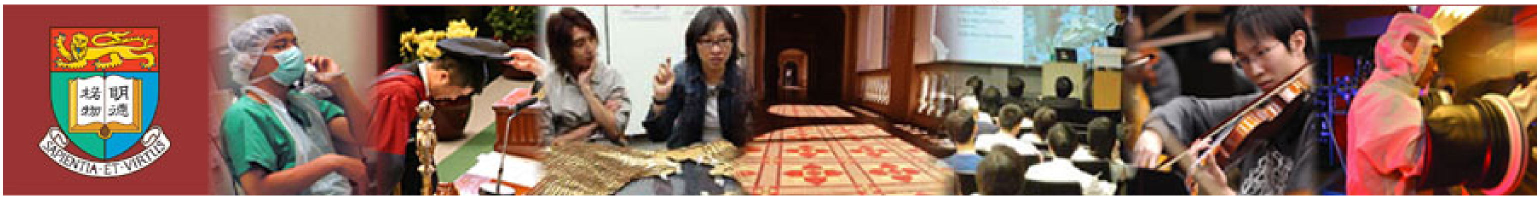
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HKU Scholars Hub

The HKU Scholars Hub The University of Hong Kong 香港大學學術庫



Title	Bayesian 3D model based human detection in crowded scenes using efficient optimization
Author(s)	Wang, L; Yung, NHC
Citation	The 2011 IEEE Workshop on Applications of Computer Vision (WACV 2011), Kona, HI., 5-7 January 2011. In Proceedings of WACV2011, 2011, p. 557-563
Issued Date	2011
URL	http://hdl.handle.net/10722/137725
Rights	Creative Commons: Attribution 3.0 Hong Kong License

BAYESIAN 3D MODEL BASED HUMAN DETECTION IN CROWDED SCENES USING EFFICIENT OPTIMIZATION

Lu Wang, Nelson H. C. Yung
Laboratory of Intelligent Transportation Systems, Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China

Abstract

In this paper, we solve the problem of human detection in crowded scenes using a Bayesian 3D model based method. Human candidates are first nominated by a head detector and a foot detector, then optimization is performed to find the best configuration of the candidates and their corresponding shape models. The solution is obtained by decomposing the mutually related candidates into un-occluded ones and occluded ones in each iteration, and then performing model matching for the un-occluded candidates. To this end, in addition to some obvious clues, we also derive a graph that depicts the inter-object relation so that unreasonable decomposition is avoided. The merit of the proposed optimization procedure is that its computational cost is similar to the greedy optimization methods while its performance is comparable to the global optimization approaches. For model matching, it is performed by employing both prior knowledge and image likelihood, where the priors include the distribution of individual shape models and the restriction on the inter-object distance in real world, and image likelihood is provided by foreground extraction and the edge information. After the model matching, a validation and rejection strategy based on minimum description length is applied to confirm the candidates that have reliable matching results. The proposed method is tested on both the publicly available Caviar dataset and a challenging dataset constructed by ourselves. The experimental results demonstrate the effectiveness of our approach.

1. Introduction

Human detection is an important task in video surveillance. It is difficult because the human objects' appearance may vary due to many factors. This task becomes even more challenging in crowded scenarios where human objects overlap with each other and therefore partial occlusion exists prevalently.

Many human detection methods can not deal with occlusion very well. e.g. the well known HOG based human detector [1]. Therefore, to detect human in more complex scenarios, substantial research works have been carried out. Most of these works use body part detectors to nominate human candidates and then perform an optimization process to select the best candidate subset as the final detection result. However, as the number of all the possible combinations of candidates is quite large, brute force search for the optimal configuration is impossible and efficient optimization method must be developed.

[2-4] use greedy methods for optimization. These methods assume an occlusion order of the candidates and decide to reject or accept a candidate sequentially from the candidate that is nearest to the camera to the farthest one. In [2], responses of part and full body detectors based on edgelet features are combined to form a joint likelihood model of human. In [3], a hierarchical part-template matching is proposed to handle partial occlusions. However, as we know that template matching is not as discriminative as learning based detectors, the greedy optimization algorithm proposed in [3] may not be sufficient to give a satisfactory detection result. To improve the efficiency of template matching, [4] proposed to use contour integration, which is calculated from integral images constructed by oriented string scans. To increase the reliability of candidate nomination, a shape context (SC) based human detector is also proposed. Combing the two detectors, the final configuration is obtained in a greedy manner.

To alleviate the demanding work required by high quality candidate nomination, global optimization methods are developed. [5] proposed to use 3D human shape models for crowd segmentation and MCMC is applied to search the solution space. Later work [6] is similar to [5] in the optimization process, except that camera calibration is estimated from the data and the shape models are learned from the data as well. [7] proposed to use EM to assign image features to human candidates, in which certainty is propagated from regions of low ambiguity to those of high ambiguity. Akin to [7], image patches are assigned to candidates using EM in [8]. The difference is that occlusion reasoning is explicitly performed in the M-step in [8] whereas [7] does not.

This paper proposes a Bayesian 3D model based approach that, given the foreground and camera parameters, segments the crowd into individuals. 3D model based approach has the advantage that it is view invariant method and it does not need, as the 2D template based method [9], to collect the large number of exemplar images to cover the shape space. We make use of the prior to model the

978-1-4244-9497-2/10/$26.00 ©2010 IEEE

distribution of an individual human object's shape, restrict inter-pedestrian overlap and require the configuration of pedestrians' locations be consist with the real world. Image likelihood is used to measure how well the detections are consistent with the foreground mask and the image's edge information. A model hierarchy is built to perform efficient model matching. Minimum description length (MDL) is applied to reject false candidates. When performing occlusion reasoning, the method is based on the following argument: generally, within a local neighborhood, true human objects have higher model matching scores than false ones and un-occluded human objects have higher model matching score than occluded ones.

The main contribution of the proposed method is a candidate optimization procedure which balances between the greedy optimization method and global optimization method. By depicting the relationship among the multiple candidates using a directed graph, candidate validation and rejection are executed in an ordered and efficient manner. In each iteration, a group of candidates are selected for model matching, which, by considering candidates that are mutually dependent, can avoid the incorrect decisions that might be made by considering only one candidate at a time [2-4]. On the other hand, as only a small portion of the candidates are considered, the computational cost is much lower than those methods which consider all the candidates at the same time [5-8].

The rest of this paper is organized as follows: Section 2 provides a theoretical formulation of the proposed method. In section 3, we introduce the implementation details of the method. In section 4, we demonstrate the performance of the system with experimental results on two datasets. Finally, we conclude the paper in section 5.

2. Problem formulation

Our goal is to find the optimal configuration of human objects, given a set of candidates, where occlusion may exist. We formulate it as a maximum a posterior (MAP) problem such that the optimal solution θ^* is given by

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

where θ consists of the number of human objects n and their corresponding models (m_i, i=1,…,n); I is the image observation. To define the prior $P(\theta)$ and the likelihood P(I|θ), we have to first define the 3D human shape model.

2.1. The 3D Human Shape Model

The 3D human shape model we propose consists of seven parts – the head (modeled by an ellipsoid), the shoulder (modeled by the upper half an ellipsoid), the torso (modeled by a cylinder), the left/right thigh and the left/right calf (each modeled by a cylinder) – as is shown in Figure 1. The dimension of the prototype model is of the average size of 50% man and 50% women presented in [10] and it is scaled linearly to generate models of different heights. To restrict the search space, ten typical leg configurations of a walking cycle are selected for model matching according to the normal walking patterns of human beings [11]. The ten configurations correspond to the five typical walking postures shown in Figure 1 and the number is doubled by differentiating left and right legs. To further consider different walking speed of human beings, the average hip and knee rotation degrees for different postures are also increased and decreased 25% respectively, by assuming local linearity in the model shape space. Therefore, the model has totally 30 postures.

In addition, the model is allowed to have 12 orientations (0°, ±30°, ±60°, ±90°, ±120°, ±150° and 180°, with 0° corresponds to human facing the camera) and four scales (corresponding to the height of 1.55m, 1.65m, 1.75m and 1.85m respectively). The head torso deviation is defined in the image space and the discretization step is set to be max(2, [W_{head}/6]), where W_{head} is the width of the head on the image.

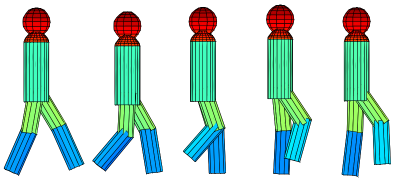


Figure 1. Illustration of various postures of the 3D human models

2.2. The Prior Distribution

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

$$P(\theta) = \prod_{i=1}^{n} P_{penal}(m_i) P_{pos}(m_i \mid m_{-i}) P_{dev}(m_i) P_{height}(m_i), \qquad (2)$$

where m_i is the shape model. The first term $P_{penal}(m_i)$ gives each model m_i in θ a penalization according to their real world position, which in fact control the allowed overlapping between models and hence avoiding n to be unreasonably large. $P_{pos}(m_i|m_{-i})$ is the prior probability of the i^{th} human object's position relative to the 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. $P_{dev}(m_i)$ and $P_{height}(m_i)$ are about the shape model itself. $P_{dev}(m_i)$ limits the head's deviation from the torso. This is used to describe our common sense that human head, when walking, tends to lean forward, but not always leans left or right and seldom lean backward. We allow the deviation but penalize the unlikely situations as they may mislead the model matching. The prior about the model height $P_{height}(m_i)$ is used to penalize very short or very tall heights. It is defined as a bell distribution such that $P_{height}(m_i)$ for the model height of 1.5 m or 1.9 m is 0.95 and for 1.7 m is 1.

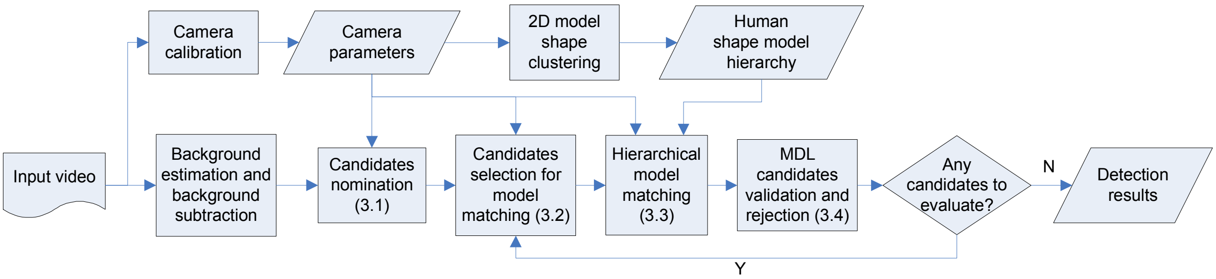

Camera calibration
Camera parameters
2D model shape clustering
Human shape model hierarchy
Input video
Background estimation and background subtraction
Candidates nomination (3.1)
Candidates selection for model matching (3.2)
Hierarchical model matching (3.3)
MDL candidates validation and rejection (3.4)
Any candidates to evaluate?
N
Y
Detection results

Figure 2. Implementation overview

2.3. Image Likelihood

Assuming the pixels are independent, the image likelihood of a solution θ is defined as

$$P(I\,|\,\theta)=\prod\nolimits_{p\in I_f}P(p\,|\,\theta)=\exp(-\sum_{p\in I_f}(1-SL_{\theta}(p))\,, \tag{3}$$

where I_f is the foreground mask, $SL_\theta(p)$ is the shape likelihood of matching the un-occluded part of the boundary of m_i with the foreground edge, if p belongs to the un-occluded part of m_i; otherwise $SL_\theta(p) = 0$, meaning that pixel p is not inside any human object models of θ.

3. Implementation

Given a video sequence, firstly, we obtain the camera parameters and perform 2D shape model clustering. Then, for each frame of the input video, we extract the foreground by the multiple adaptive thresholds method [12] where most of the shadows can be effectively removed. After that, an upper semi circle detector is used to give an exhaustive nomination of head candidates and the lower extrema detection on the mask boundary is performed to nominate foot candidates. To find the optimal configuration of the candidates, by analyzing the mask and the relationship among candidates, in each iteration of the optimization, only a group of the possible un-occluded candidates are selected for model matching, and the results are fed into a MDL based validation and rejection procedure. The iteration is repeated until all the candidates have been examined. Figure 2 gives an overview of the implementation and the following subsections explain the details.

3.1. Candidate nomination

From our observation, the most reliable feature of a human is the head. Therefore, we use an upper semi circle detector to nominate the head candidates (HCs). The applied method [13] 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) that represents the circle's center and radius. The directional filter we use is probability of boundary (*pb*) [14], which effectively removes the edge response of textures and thus reduces the number of false positive detection. The scale set of the circle detection is determined by the actual size of human heads and the camera parameters.

We also detect lower extrema (LE) on the mask boundary as foot candidates (FCs). The complementary characteristic of HCs and FCs is depicted in Figure 3. The combination of HCs and FCs forms the candidates set C_{total}.

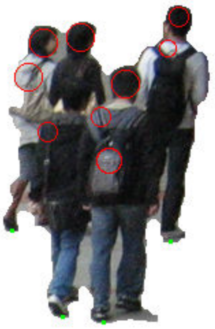


Figure 3. Illustration of the head candidates (red circles) and the foot candidates (green dots).

3.2. Candidates selection for model matching

Candidates selection is critical for the efficiency of the system: if more than enough candidates are selected and model fitted, accuracy can be guaranteed while efficiency may be sacrificed; on the other hand, if less than enough candidates are selected, accuracy may decrease. Because the candidates are mutually dependent on each other, in order to properly select the candidates for model matching, we describe candidates' occlusion relationship using a directed graph *G*, on which candidates selection will rely. To this end, for each HC, we draw its bounding polygon (BP) according to the head top position, assuming the HC has the same height as its BP. A candidate's BP is a polygon that approximately defines the maximum extents of a human model. Our BP is composed of three parts: head (rectangle, 0.3 m*0.2 m), torso (rectangle, 0.6 m*0.6 m) and lower body (trapezium, upper bottom: 0.6 m, lower bottom: 1.0 m, height: 1.2 m). The combination of the head and torso is called the upper body. For any two HCs *A* and *B*, we check their BPs: if they are not intersected or only their lower body parts intersect (Figure 4(a)), i.e. the intersection is not significant, they are not related; otherwise, if the upper body of *A*'s BP intersects with *B's* BP and A's head top is either inside B or lower than B's torso top (Figure 4(b)), then *B* is occluded by *A* and we represent this relation in *G* using an arrow starts from *A* and ends at *B*, i.e. if *A* is not matched, then *B* is not eligible for matching; otherwise, if *A* and *B*'s torsos intersect and their vertical distance of the head top is smaller than the height of a normal head (Figure 4(c)), both could be occluding the other, *A* and *B* must be matched simultaneously, i.e. if *A* is matched while *B* is not, *B* must be matched at the same time (in *G* there is a bidirectional arrow between A and B).

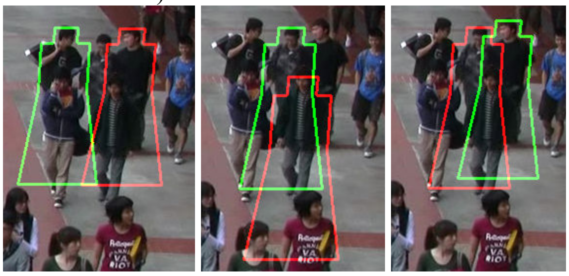

(a) (b) (c)

Figure 4. Illustration of candidates' relationship: (a) insignificant overlapping; (b) significant overlapping and significant height difference, the lower one is the higher one's preceding HC; (c) significant overlapping with similar height, the two candidates are required to match simultaneously.

For FCs, because an LE may either correspond to a true foot or just be caused by fragmented foreground, we have to identify which case the LE corresponds to. To achieve this, we define the HCs that are intersected with an FC's best fit model as the FC's related HCs, and use a dotted directional edge starts from an HC to an FC in *G* to represent this relationship. The meaning of the dotted directional edge is: once an HC related to an FC are matched, the FC and all its related HC should also be matched, ensuring that the subsequent candidate validation and rejection have enough evidence to make correct decisions. Figure 5 gives an illustration of the graph *G*.

After defining the graph *G*, we describe how to select candidates for matching in each iteration. If the bottom line of an HC's BP does not intersect with any foreground pixels, it is possible that the human object correspond to this candidate is un-occluded and we call this kind of candidates un-occluded candidates. All FCs are also taken as un-occluded candidates. In the first iteration, the candidate that is the nearest to the camera and meanwhile un-occluded is selected and model fitted. Then all the other un-occluded candidates whose BPs are intersected with the matched candidates' models are also selected and model fitted. The selection repeats until there are no more candidates satisfying the requirement.

In the following iterations, the candidates intersected with the validated models and with all their occluding candidates in *G* having been matched are selected for matching. For any un-matched un-occluded candidate, if there is one matched model whose distance to the camera is larger than its distance to the camera, this un-occluded candidate is selected for matching. (For an HC, its distance to the camera is unknown. Fortunately, for an un-occluded HC, it is reasonable to take the lowest pixel of its BP's intersection with the I_f as the position of the HC.) The candidates which are required to match simultaneously in *G* are also selected if one of them has been matched. After model matching, new candidates that satisfy the above criteria are selected. The selection ends when there are no more candidates satisfying the requirement. In case that no candidates are selected in an iteration and there are still unmatched candidates, the candidate with the smallest distance to the camera is selected.

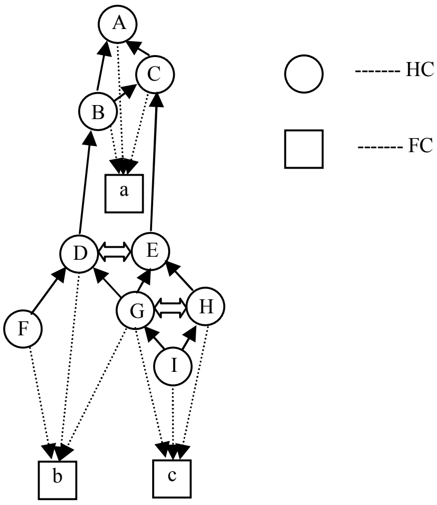


Figure 5. An illustration of the graph *G* depicting candidates' relationship.

3.3. Hierarchical model matching

Given a selected candidate c_i, if it is an HC, the head position is fixed and if it is an FC, however, as the leg postures vary a lot, the FC's position may not correspond to the exact position of the human object and we search in the vicinity of the FC to find the most proper position. The matching of the model with the image is measured by both the model's region coverage with the remained mask I_{rem} (obtained by subtracting the regions occupied by the validated models I_{occ} from I_f) and the model boundary's matching with the *pb* map of the foreground. Thus, the likelihood $L(M_j)$ is the product of the region likelihood $RL(M_j)$ and the shape likelihood $SL(M_j)$

$$L(M_j) = RL(M_j)SL(M_j)\,. \tag{4}$$

The region likelihood $RL(M_j)$ is defined as

$$LR(M_j) = c[area(M_j \cap I_{rem}) - w \cdot area(M_j \cap (1 - I_f))], \tag{5}$$

where c is a constant to ensure RL is smaller than 1; inside the brackets of (5), the minuend encourages larger area to be explained by the model while the subtrahend penalizes the model regions falling out of I; w, ranging from 0 to 1, is the penalty parameter that depends on the quality of the foreground mask: the larger the false negative rate of the foreground extraction is, the smaller w is, meaning that the region information is less reliable.

The shape likelihood SL (M_j) is defined as

$$SL(M_j) = \frac{1}{|Mb_{j,rem}|} \sum\nolimits_{k \in Mb_{j,rem}} pb(k) < \mathbf{O}_{pb}(k) \cdot \mathbf{O}_{Mb_j}(k) >,$$

$$Mb_{j,rem} = Mb_j \bigcap (1 - I_{occ}), \tag{6}$$

where Mb_j is the boundary image of model M_j, $|\cdot|$ denotes the number of non-zero pixels of the image, and $\mathbf{O}$ represents the orientation vector of the boundary point. $SL(M_j)$ is the average pb value of the un-occluded part of the model boundary weighted by the consistency between the orientation of the pb and the model boundary.

Then, according to (1)-(4), given the already validated candidates C_{val}, the matching posterior of a model M_j can be calculated as

$$P(M_j \mid C_{val}) = P_{penal}(M_j) P_{pos}(M_j \mid C_{val}) P_{dev}(M_j) P_{height}(M_j)$$
$$\exp(-RL(M_j) SL(M_j)). \tag{7}$$

and the best fit model of the candidate is defined as

$$m = \arg\max_{M_j} P(M_j \mid C_{val}) \tag{8}$$

To efficiently search a best fit model m for a candidate, we refer to [9, 15] to establish a model hierarchy for models of the same scale and with no head torso deviation. We divide the projected model shapes into seven groups according to their orientations: {0°,180°}, {30°, 150°}, {60°, 120°},{90°}, ,{-60°, -120°}, {-30°, -150°} and {-90°}. Then for each group, we construct a model shape hierarchy based on the shape dissimilarities, measured by chamfer distance between 2D model boundaries, using a agglomerative clustering method. The highest level of each hierarchy consists of two nodes. The seven hierarchies constitute the final model shape hierarchical tree with the root being empty.

When matching the highest level of hierarchical tree, all the possible scales and head torso deviations are traversed and the best matched scale and head torso deviation are adhered to that model and only the adjacent scales and deviations are searched in the matching of next level models. As in [15], at each level, the maximum matching posterior P_{max} and the minimum posterior P_{min} are computed and a threshold is selected as

$$P_\tau = P_{min} + c_\tau (P_{max} - P_{min}), \tag{9}$$

to discard the nodes that are not good enough. In our experiment, after balancing the model matching accuracy and computational cost, we set c_τ to be 0.3.

Because the prior terms $P_{height}(m)$ and $P_{dev}(m)$ also evaluates the quality of a shape model, from here on, model matching score S_m refers to

$$S_m(m) = SL(m) + \log(P_{height}(m) P_{dev}(m)). \tag{10}$$

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 nearer to the camera, indicating that the candidate is unlikely to be occluded, are preferred to be validated. We first reject the candidates that have unsatisfactory model matching quality or the candidates whose corresponding area can be better explained by other candidates, and then confirm the candidate that is less likely to be occluded by any other candidates.

a) Consider single candidate's model matching quality

For each candidate c_i that is matched in Section 3.3, if the model matching score $S_m(m_i)$ is smaller than a threshold S_T, or adding m_i into θ cannot increase the posterior $P(\theta \mid I)$, c_i is rejected.

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 as followed:

$$\begin{aligned} Sav_i &= SE_i - SE_{-i} + SM_i \\ SE_i &= area(m_{i,rem})(1 - S_m(m_i)) \\ SE_{-i} &= \max_{j,k \neq i} \sum\nolimits_{p \in m_{i,rem}} (1 - \max(S_m(m_j, p), S_m(m_k, p))) \\ SM_i &= SM_{i,0} \cdot \frac{1 - area(m_i \bigcap I_{occ})}{area(m_i)} \end{aligned}, \tag{11}$$

where $m_{i,rem}$ is m_i's intersection with I_{rem}, SE_i is the error introduced by using m_i to explain $m_{i,rem}$, SE_{-i} is the error introduced by combining other candidate models matched in the current iteration to explain $m_{i,rem}$. We limit the number of candidates for combination to be at most two because a false positive candidate can come from at most two real human objects. $S_m(m_j, p) = S_m(m_j)$ if $p \in m_j$ and $S_m(m_j, p) = 0$ otherwise. SM_i is the cost of the model after considering the portion that is occluded and $SM_{i,0}$ is the original cost of the model. If Sav_i is positive, c_i is rejected.

After rejecting the candidates that are not good enough, we examine which candidates should be validated. We perform the validation by exclusion, i.e. excluding the candidates that should not be validated and then validating the remaining ones. For any pair of intersected models, because they cannot be un-occluded at the same time, we select the one that should not be validated according to the following rules:

1). If their distance to each other is smaller than d_{min} (the minimum permissible distance for two human objects), or their overlapping area is larger than 90% of the area of the smaller model, or they are of left-right relation, any one could be un-occluded. Therefore, we compare their posterior first, if one's posterior is significantly larger than the other (the parameter that indicates "significantly larger" is learned through experiments and is fixed at 1.35 for all the tested images), the one with lower posterior is not validated; otherwise, we compare their shape matching score calculated by (10) and the one with lower score is not validated.

2). Otherwise, the one that is nearer to the camera should be un-occluded and the one that is farther away from the camera is not validated.

After the validation, the validated candidates are then added to θ, and their covered regions are deleted from I_{rem} and added to the occupancy map I_{occ}. The whole optimization procedure is summarized as followed.

```
Algorithm: optimization algorithm
Given the candidate nomination $C_{total}$ and the foreground mask $I_f$,
initialize $\theta = \emptyset$, $I_{occ}$ as empty (black image), $I_{rem} = I_f$, the validated candidates set $C_{val} = \emptyset$, the rejected candidates set $C_{rej} = \emptyset$, and the posterior as $P(\theta \mid I) = \exp(-area(I_f))$.
Build the candidates' relation graph $G$.
while $C_{val} \cup C_{rej} \neq C_{total}$
    1. Select the possible currently un-occluded candidates. (3.2)
    2. For each selected candidate, perform hierarchical model matching. (3.3)
    3. Reject and validate these matched candidates and update $C_{rej,}$ $C_{val}$, $\theta$, $I_{rem}$, and $I_{occ}$. (3.4)
end
return $\theta$.
```

4. Experimental result

We evaluate the proposed method using two datasets: the first one is the Caviar benchmark dataset [16] and the second one is an outdoor scene video taken in our campus. Foreground of the Caviar dataset is more fragmented than the video taken by us and hence less reliable. Therefore, the parameter w in (5) is set to be 0 for the Caviar data and 0.8 for our campus video; all the other parameters are set the same.

The evaluation is based on the following criteria: a) a correct detection is a detection DT that has a one-to-one correspondence GT in the ground truth human objects and satisfying

$$\text{Overlap}(GT,DT) = \frac{\text{area}(GT \cap DT)}{\text{area}(GT \cup DT)} > 0.5, \qquad (12)$$

and b) human objects having less than 50% of the bodies inside the images are not evaluated; c) sitting and scene occluded (more than 20% occluded) human objects are not evaluated; d) human objects staying in the scene for a relatively long time without significant movements are not evaluated and considered as scene objects.

4.1. Detection results on the Caviar dataset

We evaluated the proposed method on the sequence OneStopMoveEnter1Cor (1590 frames with image resolution being 384 × 288) of the Caviar dataset. To compare the proposed method with previous works [3], in which evaluation is done for 200 selected frames of this sequence, and [4], in which evaluation is done for frames 800-1000, we also evaluated our method for frames 800-1000. The ROC curves for different methods are plotted in Figure 6. As can be seen from the ROC curves, the proposed method has a detection rate around 99% with tolerable number of false alarms. Figure 7 illustrates some detection results.

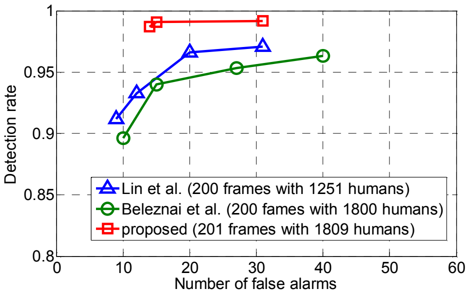


Figure 6. ROC curves of evaluation on a subset of the Caviar dataset.

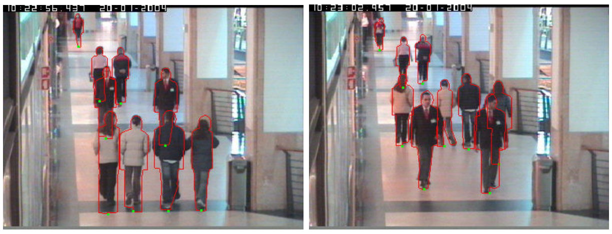


Figure 7. Illustration of detection results on Caviar dataset.

4.2. Detection results on campus dataset

The campus dataset consists of 50 minutes of video taken at 25 f/s and a resolution of 1280×720. The view is deep and wide, resulting in substantial scale changes (with the width of a normal human object ranging from 10 pixels to 70 pixels). In addition, on the right hand side of the scene, the illumination is weak and the background is dark.

Due to the large number of frames, we sub-sampled the frames to 2.5 f/s, obtaining 7500 frames, on which the proposed method was tested. However, 7500 frames still represent a sizeable evaluation task. As such, we manually selected 500 frames (containing 7116 humans) where occlusion occurs frequently and the number of humans is relatively large. Figure 8 illustrates some detection results. The detection rate achieved is **90.9%** when the false positive rate is **1.53%**. Among the errors, missed detections mainly come from low foreground/background contrast and low resolution, whereas false alarms usually appear in texture rich regions.

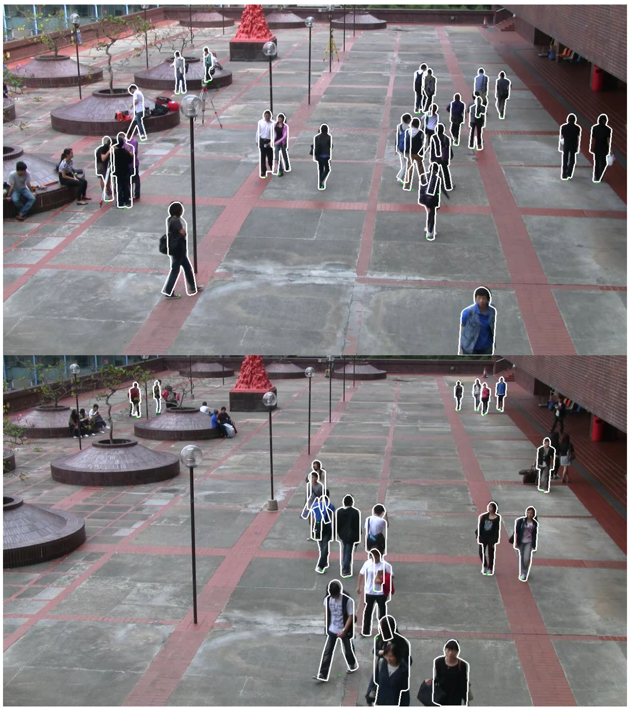


Figure 8. Illustration of detection results on the campus dataset.

4.3. Computational cost analysis

By randomly selecting 2263 candidates and counting the number of times they are selected for matching during the optimization process, we obtained the result as shown in TABLE I. It can be seen that 62.1% of the candidates are just visited for once and more than 87% of the candidates are visited no more than twice. The average visited times is 1.57 for these candidates. This result demonstrates that our method does not cost much more than those greedy methods in which each candidate is visited for only once.

TABLE I. NUMBER OF TIMES VISITED FOR 2263 CANDIDATES

Times visited	1	2	3	4	5	6	7	>=8
No. of candidates	1405	569	188	63	26	10	2	0
%	62.1	25.1	8.3	2.8	1.1	0.4	0.1	0

5. Conclusion

A Bayesian approach for human detection in crowd scenarios has been proposed in this paper. Foreground and edges are used to provide image evidence for the inference. Knowledge priors about human shape distribution and inter-human minimum distance limitation are enforced during the model matching process. The solution is obtained in a way that balances the computational cost and the performance. Results on various data show the effectiveness of the proposed method.

To improve the performance, the most important future work is to combine the detection results across consecutive frames, which can resolve the ambiguities of a single frame, to obtain a more reliable detection performance.

Acknowledgment

This work is supported in part by 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.

References

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

[2] 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.

[3] Z. Lin, *et al.*, "Hierarchical part-template matching for human detection and segmentation," in *ICCV*, 2007, pp. 1-8.

[4] C. Beleznai and H. Bischof, "Fast human detection in crowded scenes by contour integration and local shape estimation," in *CVPR* 2009, pp. 2246-2253.

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

[6] W. Ge and R. Collins, "Marked point processes for crowd counting," in *CVPR* 2009, pp. 2913-2920.

[7] J. Rittscher, *et al.*, "Simultaneous estimation of segmentation and shape," in *CVPR* 2005, pp. 486-493.

[8] P. Tu, *et al.*, "Unified crowd segmentaton," in *ECCV* 2008.

[9] D. Gavrila, "Pedestrian Detection from a Moving Vehicle " in *ECCV* 2000, pp. 37-49.

[10] A. R. Tilley and H. D. Associates, *The Measure of Man and Woman: Human Factors in Design*, 1993.

[11] M. P. Murray, *et al.*, "Walking patterns of normal men," *Journal of Bone and Joint Surgery,* vol. 46-A, pp. 335-360, 1964.

[12] L. Wang and N. H. C. Yung, "Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation," *TITS,* vol. 11(1), pp. 40-51, 2010.

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

[14] D. R. Martin, *et al.*, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *TPAMI,* vol. 26(5), pp. 530-549, 2004.

[15] B. Stenger, *et al.*, "Model-based hand tracking using a hierarchical Bayesian filter," *TPAMI,* vol. 28(9), pp. 1372-1384, 2006.

[16] *Caviar Test Case Scenarios*. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/