2011

Detecting humans under occlusion using variational mean field method

Duc Thanh Nguyen
University of Wollongong, dtn156@uow.edu.au

Philip Ogunbona
University of Wollongong, philipo@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Detecting humans under occlusion using variational mean field method

Abstract

This paper proposes a human detection method using variational mean field approximation for occlusion reasoning. In the method, parts of human objects are detected individually using template matching. Initial detection hypotheses with spatial layout information are represented in a graphical model and refined through a Bayesian estimation. In this paper, mean field method is employed for such an estimation. The proposed method was evaluated on the popular CAVIAR-INRIA dataset. Experimental results show that the proposed algorithm is able to detect humans in severe occlusion within reasonable processing time.

Keywords

method, field, mean, humans, variational, detecting, occlusion, under

Disciplines

Physical Sciences and Mathematics

Publication Details

Nguyen, D., Ogunbona, P. & Li, W. (2011). Detecting humans under occlusion using variational mean field method. 18th IEEE International Conference on Image Processing, ICIP 2011 (pp. 2049-2052). USA: IEEE.

DETECTING HUMANS UNDER OCCLUSION USING VARIATIONAL MEAN FIELD METHOD

Duc Thanh Nguyen, Philip Ogunbona, and Wanqing Li

Advanced Multimedia Research Lab, ICT Research Institute
School of Computer Science and Software Engineering
University of Wollongong, Australia

ABSTRACT

This paper proposes a human detection method using variational mean field approximation for occlusion reasoning. In the method, parts of human objects are detected individually using template matching. Initial detection hypotheses with spatial layout information are represented in a graphical model and refined through a Bayesian estimation. In this paper, mean field method is employed for such an estimation. The proposed method was evaluated on the popular CAVIAR-INRIA dataset. Experimental results show that the proposed algorithm is able to detect humans in severe occlusion within reasonable processing time.

Index Terms— Human detection, occlusion reasoning, mean field method

1. INTRODUCTION

Human detection from still images and videos is a crucial step in human motion analysis that is currently receiving much attention in computer vision area. The challenges of this task arise from many factors, including, the complexity of the background, the variation of human appearance, postures-viewpoints, and occlusion. In recent years, many human detection algorithms employing different feature descriptors have been developed. For example, robust features have been proposed to encode human appearance. In [1], simple curves and segments called 'edgelets' were employed to describe the body parts of human object. A well-known feature, namely, histogram of oriented gradients (HOG) was introduced in [2]. Recently, local binary patterns (LBP) were employed to describe the human body [3].

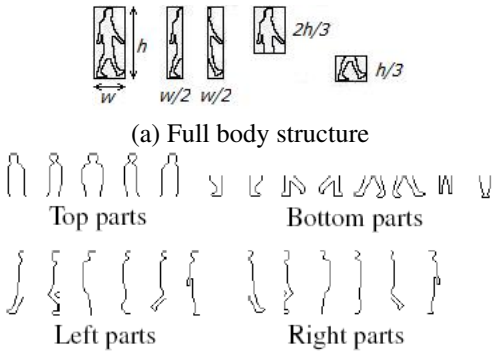
To overcome the variation of human postures and viewpoints, template matching is often used. Edge templates are employed to represent shapes of various postures and viewpoints of the full human body [4, 5] or body parts [6, 7]. Another approach is the use of the implicit shape model (ISM) [8] to represent the spatial constraint between the body parts.

One of the most difficult challenges in human detection is occlusion. A number of methods addressing occlusion problem have been proposed in the literature. In general, these methods start with detecting the body parts and then infer the occlusion using some reasoning algorithms. For example, Zhao et al. [9] formulated the inference process as an optimisation problem and Markov chain Monte Carlo (MCMC) was applied to find the optimal solution. Similar to [9], the problem was formulated as an optimization task in [1, 6, 10], but a greedy-based strategy was used in the solution. In [11], a logic based reasoning framework was proposed for the occlusion inference. The framework used a number of logical rules based on the response of each individual part detector and the geometric constraints between detected parts.

In this paper, we address the problem of detecting humans under occlusion by adopting a template matching-based approach to provide an initial set of human candidates. We then refine this set by formulating the problem as a Bayesian estimation solved using the variational mean field method. The remainder of this paper is organised as follows. Section 2 briefly presents a template matching-based human detection method. The occlusion problem is formulated and a solution is proposed in section 3. Experimental results are shown in section 4. Section 5 concludes the paper and provides some remarks.

2. HUMAN DETECTION USING SHAPE MATCHING

A well-known advantage of template matching approach is that templates can be used to describe humans in various postures and viewpoints. In addition, part-based detection is appropriate for detecting humans under occlusion since not all body parts are fully visible in this situation. In this paper human detection is performed using part-based template matching. Fig. 1 shows the part-based template model [7] used in the paper with 5, 8, 6, and 6 templates representing the top part, bottom, left, and right parts respectively (readers are referred to [7] for more details in creating the part template model). As can be seen in Fig. 1, the number of templates matched is $5 + 8 + 6 + 6 = 25$ (templates) to cover up to $5 \times 8 \times 6 \times 6 = 1440$ postures. Compared with the full body detection approach, this provides an advantage since the matching is performed on a small set of templates to cover a



(a) Full body structure

(b) Part templates

Fig. 1. Part template model used in this paper [7].

variety of human postures.

Given a part template model as presented in Fig. 1, each part $i \in \{top, bottom, left, right\}$ can be represented by a set of part templates S_i. The detection algorithm is performed by scanning an input image with a detection window W at various scales and positions. A hypothesis about the presence of human can be generated by verifying the matching score $C(W)$ with a threshold θ as,

$$C(W) = \frac{\sum_i v_i[1 - D(p_i^*, W_i)]}{\sum_i v_i} < \theta \qquad (1)$$

where $i \in \{top, bottom, left, right\}$, $v_i \in \{0, 1\}$ indicating the presence/absence of part i, and $D(p_i^*, W_i) \in [0, 1]$ is the spatial-orientation Chamfer distance [5] between the image region W_i of the window W corresponding to part i and its best matching template p_i^* computed as,

$$p_i^* = \arg\min_{p \in S_i} D(p, W_i) \qquad (2)$$

In (1), v_i can be determined as,

$$v_i = \begin{cases} 1, & \text{if } occ(i) < \delta \\ 0, & \text{otherwise} \end{cases}$$

where $occ(i)$ indicates the ratio of the area of part i occluded by other detection hypotheses and δ represents the degree of occlusion accepted by the method.

Essentially, $C(W)$ is the average of the partial matching scores of parts appearing on the detection window W. Now, in the first stage of detection, we do not know which parts are occluded. Therefore, we assume that all parts are fully observable. This can be done by setting the threshold θ to an appropriate value so that true detections are not missed and an initial set of human candidates may contain many false alarms. However, this set is then refined using the occlusion reasoning method presented in the next section.

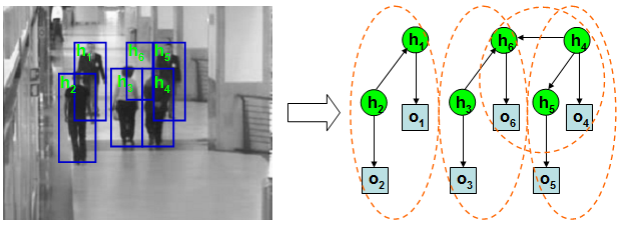


Fig. 2. Left: initial detection hypotheses, Right: the graphical model in which ellipses represent local groups.

3. OCCLUSION SOLVING USING MEAN FIELD

Given an input image I and an initial set of hypotheses about the presence of humans, $H = \{h_1, h_2, ..., h_N\}$ with corresponding image observation data $O = \{o_1, o_2, ..., o_N\}$ detected by part-based template matching method presented in section 2. Consider each hypothesis h_i and its image observation o_i as a hidden and observed node of a graph G. $h_i, i \in \{1, ..., N\}$ are binary random variables which take values in $\{0, 1\}$ to indicate false positive and true positive respectively. If h_i is occluded by h_j, there is an edge from h_j to h_i. We assume that if h_i is occluded by h_j then foot position of h_j must be higher than that of h_i. This assumption is reasonable for most surveillance systems where the images/videos are captured by a camera looking down to the ground plane. Fig. 2 shows an example of the graphical model G. As can be seen, G can be considered as a Bayesian network in which h_i are state variables. The initial set of hypotheses H may contain false alarms which have been generated without considering occlusion reasoning. Refining this set corresponds to making inference on appropriate values of hypotheses h_i in estimating the marginal probability of the observed data:

$$\log P(O) = \log \sum_H P(O|H)P(H) \qquad (3)$$

where $P(H)$ is the prior and $P(O|H)$ is the likelihood to obtain the observed data O given hypotheses H.

Since each h_i takes a binary value, a brute force estimation of (3) would require $O(2^N)$ operations. Therefore, instead of directly estimating $P(O)$ using (3), we approximate it by finding a variational distribution Q which is also an approximate of the posterior $P(H|O)$. As presented in [12], this task can be transformed into an optimisation problem and solved using variational mean field method. In particular, an objective function is defined as,

$$J(Q) = \log P(O) - KL(Q(H)||P(H|O)) \qquad (4)$$

where KL is the Kullback-Leibler divergence of two distributions which is computed as,

$$KL(Q(H)||P(H|O)) = \sum_H Q(H) \log \frac{Q(H)}{P(H|O)} \qquad (5)$$

Substituting (5) into (4), $J(Q)$ can be rewritten as,

$$\begin{aligned} J(Q) &= \log P(O) - \sum_H Q(H) \log \frac{Q(H)P(O)}{P(H,O)} \\ &= -\sum_H Q(H) \log \frac{Q(H)}{P(H,O)} \\ &= -\sum_H Q(H) \log Q(H) + \sum_H Q(H) \log P(H,O) \\ &= \mathcal{H}(Q) + E_Q\{\log P(H,O)\} \end{aligned} \tag{6}$$

where $\mathcal{H}(Q)$ is the entropy of the variational distribution Q, and $E_Q\{\cdot\}$ represents the expectation with regard to Q.

Since the KL-divergence is nonnegative, maximising the lower bound $J(Q)$ with regard to Q will give us an approximate $J(Q^*)$ of $\log P(O)$ and Q^* of the posterior $P(H|O)$. In addition, approximation of $\log P(O)$ relates to finding an appropriate variational distribution $Q(H)$. In this paper, the simplest selection of variational distributions which assumes that all hidden variables are independent of each other is adopted. In particular, we assume,

$$Q(H) = \prod_{i=1}^{N} Q_i(h_i) \tag{7}$$

Thus, the entropy $\mathcal{H}(Q)$ can be rewritten as,

$$\mathcal{H}(Q) = \sum_{i=1}^{N} \mathcal{H}(Q_i) \tag{8}$$

where $\mathcal{H}(Q_i)$ is the entropy of the variational component Q_i.

Since $Q(H)$ is fully factorised, $J(Q)$ can be optimised with regard to each individual component Q_i at a time. Thus, $J(Q)$ can be estimated by updating the i-th component while other components remain unchanged, i.e.,

$$J(Q) = \text{const.} + \mathcal{H}(Q_i) + \sum_{h_i} Q_i(h_i) E_Q\{\log P(H,O)|h_i\} \tag{9}$$

where $E_Q\{\cdot|h_i\}$ is the conditional expectation with respect to the variational distribution Q given h_i.

As presented in [12], maximising $J(Q)$ can be obtained by computing Gibbs distributions of $Q_i(h_i)$:

$$Q_i(h_i) \leftarrow \frac{1}{Z_i} e^{E_Q\{\log P(H,O)|h_i\}} \tag{10}$$

where Z_i is the normalisation factor computed as,

$$Z_i = \sum_{h_i} e^{E_Q\{\log P(H,O)|h_i\}} \tag{11}$$

Update equations (10) and (11) will be invoked iteratively to increase the objective function $J(Q)$. It can be seen that the update of $E_Q\{\cdot|h_i\}$ depends only on h_i and hypotheses occluded by h_i. Thus, $E_Q\{\cdot|h_i\}$ can be factorised over *local groups* $\mathcal{C}(h_i) = \{(h_i, o_i, h_j, o_j)\}$ where h_j is occluded by h_i (see Fig. 2 for an example of local groups). In particular, the update can be performed as,

$$E_Q\{\log P(H,O)|h_i\} \leftarrow \sum_{c \in \mathcal{C}(h_i)} \sum_{h_j \in \{0,1\}} Q_j(h_j) \log \psi(c) \tag{12}$$

where $\psi(c)$ is the potential function of the *local group* c and can be computed conventionally as in a Bayesian network:

$$\psi(c) \equiv P(h_i, o_i, h_j, o_j) = P(o_i|h_i)P(h_j|h_i)P(o_j|h_j)P(h_i) \tag{13}$$

where we define $P(o_i|h_i) = C(o_i)$ (since h_i is assumed to represent a human) computed using (1). In addition, we set $P(h_i)$ and $P(h_j|h_i)$ to positive constants α and β respectively. If h_i does not occlude any other hypotheses, $\psi(c)$ will be simplified to $P(o_i|h_i)$. Finally, if $Q_i(h_i = 1) \geq Q_i(h_i = 0)$, h_i is set to 1, i.e. true detection, and $P(o_j|h_i)$ is re-evaluated using (1) accordingly with current setting of h_i. When the optimal Q^* is found, the corresponding subset of hypotheses $h_i = 1$ will be determined. This subset provides the final detection results.

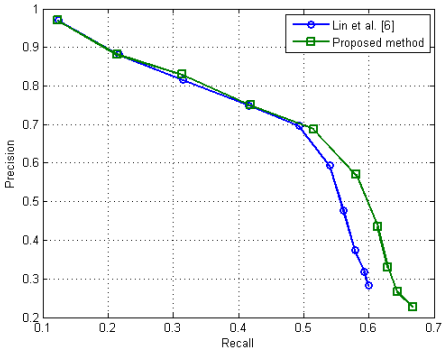


Fig. 3. PR Curves of the proposed method and Lin's method [6] on the *OneStopMoveEnter1cor* sequence.

4. EXPERIMENTAL RESULTS

There are a number of parameters used in this paper including the rejection threshold θ in (1), occlusion degree δ, and two constants α and β assigned to $P(h_i)$ and $P(h_j|h_i)$. In our implementation, without any knowledge about the occlusion cases, we set δ to 0.65. We also assume that h_i is distributed uniformly, i.e. $P(h_i = 0) = P(h_i = 1) = \alpha = 0.5$ and $P(h_j|h_i) = \beta = 0.5$ for all $h_i, h_j \in \{0, 1\}$. In addition, θ is varied to represent the trade-off between true detections and false alarms.

The proposed method was evaluated on the *OneStopMoveEnter1cor* sequence of the CAVIAR-INRIA dataset[1]. This sequence consists of 1590 images of 384×288 pixels. A subset of 200 images (800th-1000th frame) with 1614 annotated humans was created for testing the method. We used the PR (Precision-Recall) measure to evaluate the detection performance. Fig. 3 shows the PR curve of the proposed method. Some detection results are shown in Fig. 4. As can be seen from Fig. 4, occlusion reasoning could reduce false alarms while retaining correct detections. However, false detections are still present in the image. This is due to the weakness of the template matching-based detector. Employing more sophisticated detectors, e.g. HOG detector [2], might improve the detection performance. However, development of robust human detectors is not the main focus of this paper.

As presented in section 3, the occlusion inference is performed iteratively to maximise the objective function $J(Q)$ defined in (4); at each iteration all hypotheses are verified and corresponding update equations are invoked. Therefore, to evaluate the efficiency of the proposed method, we count the total number of iterations performed to maximise $J(Q)$ as well as the real processing time required per frame. Through experiments, we have found that the average loop time on over 200 images of the *OneStopMoveEnter1cor* sequence is about 2.1 and each frame can be processed in 1.25 second.

In addition to evaluation, we compared our method with other algorithms. In particular, the heuristic method proposed by Lin et al. [6] was selected for this purpose. To obtain a fair comparison, we used the same template matching-based human detector with the same template model and then re-implemented the heuristic-based occlusion reasoning in the work of Lin et al. [6]. Experimental results have shown that our method provides an improved performance. In particular, the improvement can be seen clearly at high recalls ≥ 0.5. For example, at a recall of ≈ 0.57, compared with Lin's method, we could increase the precision by $\approx 20\%$. The PR curves of both methods are presented in Fig. 3.

5. CONCLUSIONS

This paper proposes a human detection method under occlusion using the variational mean field. The prosed method is performed in a two-stage framework. Body parts are first detected using template matching to form an initial set of human candidates. This set is then refined using occlusion reasoning formulated as a Bayesian estimation. In this paper, we propose the use of variational mean field method to approximate such estimation. The proposed method was evaluated on the CAVIAR-INRIA dataset. The results indicate the robustness and efficiency of the proposed method in detecting humans under occlusion.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

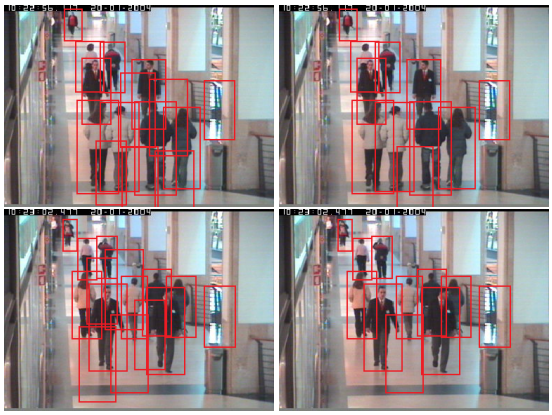


Fig. 4. Illustration of occlusion reasoning. Left: initial detection results. Right: final results obtained using reasoning.

6. REFERENCES

[1] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *ICCV*, 2005.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[3] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *CVPR*, 2008.

[4] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *PAMI*, vol. 29, no. 8, pp. 1408–1421, 2007.

[5] D. T. Nguyen, W. Li, and P. Ogunbona, "An improved template matching method for object detection," in *ACCV*, 2009.

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

[7] D. T. Nguyen, W. Li, and P. Ogunbona, "A part-based template matching method for multi-view human detection," in *IVCNZ*, 2009.

[8] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *CVPR*, 2005.

[9] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *PAMI*, vol. 30, no. 7, pp. 1198–1211, 2008.

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

[11] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis, "Bilattice-based logical reasoning for human detection," in *CVPR*, 2007.

[12] T. S. Jaakkola, "Tutorial on variational approximation methods," Tech. Rep., MIT Artificial Intelligence Laboratory, 2000.