2013

Inter-occlusion reasoning for human detection based on variational mean field

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

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

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

Publication Details

Nguyen, D., Li, W. & Ogunbona, P. O. (2013). Inter-occlusion reasoning for human detection based on variational mean field. Neurocomputing, 110 51-61.

Inter-occlusion reasoning for human detection based on variational mean field

Abstract

Detecting multiple humans in crowded scenes is challenging because the humans are often partially or even totally occluded by each other. In this paper, we propose a novel algorithm for partial inter-occlusion reasoning in human detection based on variational mean field theory. The proposed algorithm can be integrated with various part-based human detectors using different types of features, object representations, and classifiers. The algorithm takes as the input an initial set of possible human objects (hypotheses) detected using a part-based human detector. Each hypothesis is decomposed into a number of parts and the occlusion status of each part is inferred by the proposed algorithm. Specifically, initial detections (hypotheses) with spatial layout information are represented in a graphical model and the inference is formulated as an estimation of the marginal probability of the observed data in a Bayesian network. The variational mean field theory is employed as an effective estimation technique. The proposed method was evaluated on popular datasets including CAVIAR, iLIDS, and INRIA. Experimental results have shown that the proposed algorithm is not only able to detect humans under severe occlusion but also enhance the detection performance when there is no occlusion.

Keywords

inter, variational, detection, reasoning, field, occlusion, mean, human

Disciplines

Engineering | Science and Technology Studies

Publication Details

Nguyen, D., Li, W. & Ogunbona, P. O. (2013). Inter-occlusion reasoning for human detection based on variational mean field. Neurocomputing, 110 51-61.

Inter-Occlusion Reasoning for Human Detection Based on Variational Mean Field

Duc Thanh Nguyen[a,*], Wanqing Li[a], Philip O. Ogunbona[a]

[a]*School of Computer Science and Software Engineering,*
University of Wollongong, NSW 2522
Australia

Abstract

Detecting multiple humans in crowded scenes is challenging because the humans are often partially or even totally occluded by each other. In this paper, we propose a novel algorithm for partial inter-occlusion reasoning in human detection based on variational mean field theory. The proposed algorithm can be integrated with various part-based human detectors using different types of features, object representations, and classifiers. The algorithm takes as the input an initial set of possible human objects (hypotheses) detected using a part-based human detector. Each hypothesis is decomposed into a number of parts and the occlusion status of each part is inferred by the proposed algorithm. Specifically, initial detections (hypotheses) with spatial layout information are represented in a graphical model and the inference is formulated as an estimation of the marginal probability of the observed data in a Bayesian network. The variational mean field theory is employed as an effective estimation technique. The proposed method was evaluated on the popular datasets including CAVIAR, iLIDS, and INRIA. Experimental results have shown that the proposed algorithm is not only able to detect humans under severe occlusion but also enhance the detection performance when there is no occlusion.

Keywords:
Occlusion reasoning, non-redundant local binary pattern, mean field method

1. Introduction

In recent years, human detection has received much attention in applications such as video surveillance, motion analysis, and driving assistant systems [1, 2]. However, the challenges of this task are well known due to the complexity of the background, the variations of human appearance, postures and viewpoints. The problem becomes extremely difficult when detecting multiple humans in highly crowded scenes where humans can be severely occluded. In general, an object is considered occluded if it is not fully perceivable or observable. There are three types of occlusion: self-occlusion, non-object-occlusion, and inter-object occlusion.

Self-occlusion is the case in which some parts of the object are occluded by other parts of the same object. This type of occlusion is mainly caused by the variation of the camera's viewpoint or pose of the human object. Figure 1(a) shows examples of self-occlusion in which the arms of the human object are

[*] Corresponding Author. Tel: +61 2 4221 3103, Fax: +61 2 4221 4170
Email addresses: dtn156@uowmail.edu.au (Duc Thanh Nguyen), wanqing@uow.edu.au (Wanqing Li), philipo@uow.edu.au (Philip O. Ogunbona)

occluded and un-occluded by the torso; and one leg is occluded and un-occluded by another leg. It can be seen that at a certain level of the variation of posture and viewpoint, self-occlusion can be solved by employing a multi-view object descriptor, e.g. [3]. Furthermore, it is not necessary that all parts of a human must be visible so that the human can be detected. Some parts, e.g. the arms and hands, are subject to much variability because of articulation and often not seen clearly in low resolution images. Thus, they are not usually modelled separately in part-based human detection algorithms, e.g. [4, 5].

Inter-object-occlusion occurs when an object of interest is occluded by other objects. Such occlusion can be further divided into type-I and type-II inter-object occlusion. Type-I inter-object occlusion occurs when an object of interest is occluded by another object of the same type (e.g. humans occluding other humans). Type-II inter-object occlusion occurs when an object of interest is occluded by objects that are not of interest in the specific application. For human detection, the type-I inter-object occlusion, as shown in Figure 1(c), is often found in video surveillance of a dense crowd. In these applications, the camera is often set up to look down towards the ground plane where, as shown in Figure 2, inter-object occlusion can occur when a human object is blocked by another human object standing between the first human object and the camera. Examples of type-II inter-object occlusion can be seen in Figure 1(b), wherein the humans are occluded by a car, chair, table, flag.

This paper focuses on the type-I inter-object occlusion, referred to as inter-object occlusion hereafter for brevity, and proposes a method for modeling the inter-object occlusion and determining the occlusion status of the parts of human objects to improve the detection rate in a crowded scene. Specifically, given an input image, an initial set of human hypotheses is formulated using a part-based human detector. The initial hypotheses may contain false positives which have been generated without considering the occlusion status of parts. In this paper, hypotheses and their spatial relationships are represented as a graphical model and the problem of occlusion reasoning is formulated as estimating a marginal probability of the observed data. This task corresponds to making inference on appropriate status of hypotheses to explain the observation. For efficient computation, the variational mean field method is used to estimate the marginal probability. The proposed occlusion reasoning algorithm was evaluated on commonly used datasets including CAVIAR, iLDS, and INRIA. Experimental results have verified the effectiveness and robustness of the proposed algorithm in detecting multiple and partially occluded humans.

The remainder of this paper is organised as follows. In Section 2, we briefly review the related works on human detection as well as occlusion reasoning. Section 3 describes a part-based human detector which is used to obtain a set of initial human hypotheses and provides partial detection results for later occlusion analysis. The problem of occlusion reasoning is formulated in Section 4 and a variation mean field algorithm is presented in Section 5. Experimental results along with comparative analysis are shown in section 6. Section 7 discusses some aspects of the proposed method and concludes the paper with remarks.

2. Related Work

Generally speaking, existing object detection methods can be categorised as either global or local approach. Global methods focus on detection of a full object using a full object template matcher [6, 7]. Local methods on the other hand detect objects by locating parts constituting the objects [8, 4]. Compared with global detection, local detection has advantage of being able to detect objects with high articulation such as human bodies and to cope with the problem of occlusion.

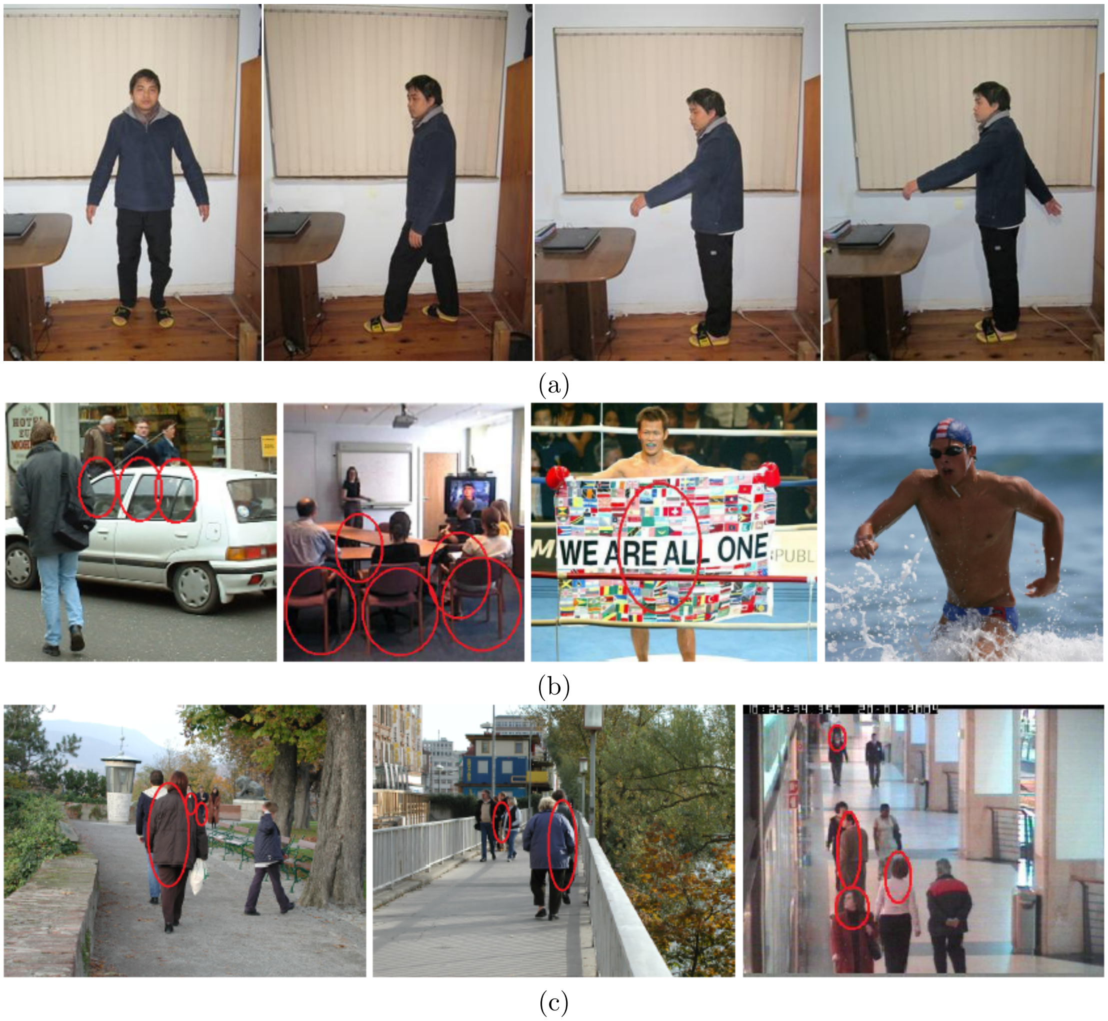


Figure 1: Illustration of different types of occlusion: (a) self-occlusion, (b) non-object-occlusion and (c) inter-occlusion, where the occluded areas are highlighted by ellipses.

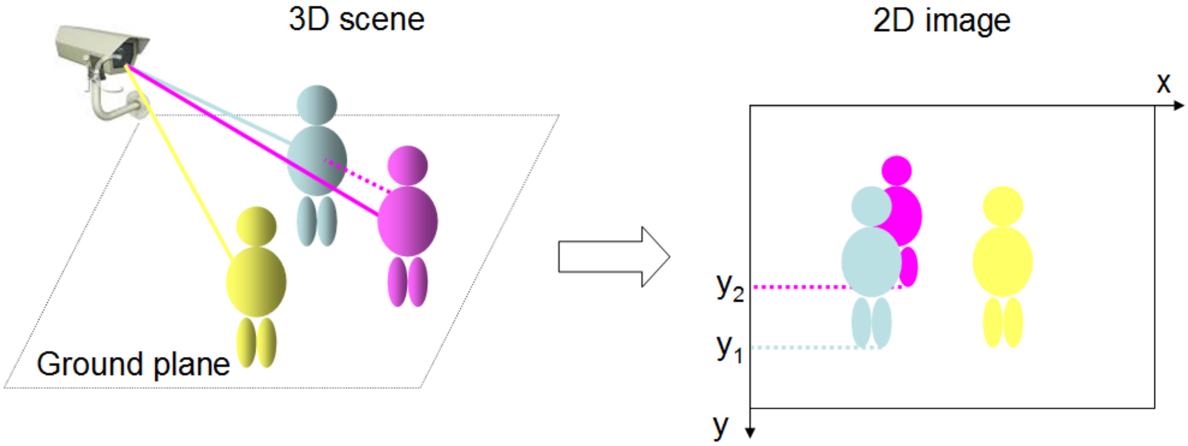


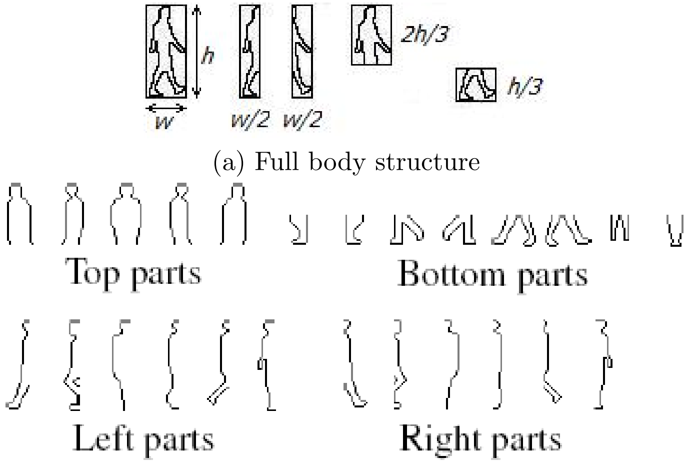
Figure 2: Ground plane in 3D scene and the corresponding 2D captured image.

In the history of object detection, many robust features have been proposed to describe the full object or object parts [1]. These features can be obtained from the low-level information of the object images such as edge, texture, or colour. To encode the shape of an object or object parts, edge-based features have often been used. For example, edge contours were predefined through templates in [6, 9, 10], edgel elements, namely as edgelets in [4, 3], or learned from samples in [11, 12]. In [7], histogram of oriented gradients (HOG) was proposed. HOG has then received much attention with various extensions, e.g. [13, 14]. In [15], the covariance matrices of the spatial location, intensity derivatives, and edge orientations were used. Since the covariance matrices do not lie on a vector space, they were classified on Riemannian manifolds. In [16], Gao et al. proposed the Adaptive Contour Feature (ACF) constructed based on the edge magnitudes and orientations in a scale space.

In addition to edge-based features, appearance features describing the texture [8, 17, 18, 19] or colour [20] have also been explored. For example, Mohan et al. [8] used Haar wavelets to describe the object texture by encoding the oriented and intensity differences between adjacent regions. Extended from Haar wavelets, Viola et al. [17] proposed rectangular features with various configurations. These features were then applied to encode movement patterns of pedestrians in [21]. Grayscale (intensity) patterns were employed to represent the local appearance of the object parts in [18]. Recently, local binary patterns (LBP) originally proposed for texture classification [22] were employed in [19, 23]. In [20], a soft segmentation based on the foreground/background colour was performed by a Fisher discriminant. HOG was then applied on the segmented image to compute the so-called CHOG feature.

One of the most difficult challenges in human detection is occlusion. A number of methods addressing the occlusion problem have been proposed in the literature. In general, these methods can be categorised as window-based or context-based approaches. The window-based approach [23, 24] has been more successful for situations where the occlusion of human objects is caused by non-human objects and the problem is solved within each image window. For example, in [23], responses of a holistic linear SVM classifier to HOG features computed on blocks in the detection window were used to construct an occlusion map for each human hypothesis. The responses with nearby values on the occlusion map were merged and segmented into regions using mean-shift algorithm [25]. Regions of mostly negative responses were inferred as occluded regions while positive regions (implied as non-occluded regions) were classified using sub-classifiers. A disadvantage of this method is that it is not applicable when other types of object representation or classifiers (not linear SVM) are employed. In [24], motion (optical flow) and depth (stereo vision) cues were incorporated in identifying non-occluded regions. This is motivated by the observation that occluded regions often cause significant discontinuity in motion flows while occluding obstacles are closers (in depth) to the camera than occluded objects. This method requires the motion and depth information and thus is not applicable to detecting occluded humans in static images. Some other methods, e.g. [26], experimentally showed that they could deal with the occlusion problem. However, there is no explicit mechanism to deal with occlusion proposed in those methods.

For context-based methods, they often start from the detection of body parts and are then followed by inferring possible inter-object occlusion through a reasoning algorithm. For example, Zhao et al. [27] formulated the inference process as an optimisation problem and Markov chain Monte Carlo (MCMC) was applied to find the optimal solution. Similarly, the problem was formulated as the maximisation of a joint likelihood of human hypotheses in [4, 28, 5, 29, 30, 31] and a greedy-based inference algorithm was used to obtain the optimal solution. In [4, 28, 5, 29, 30], human hypotheses were sorted in descending order of vertical coordinate (with respect to the image coordinate) and the optimisation was performed



(a) Full body structure

(b) Part templates

Figure 3: Part template model.

by verifying whether each hypothesis should be added (e.g. [5, 29, 30]) or removed (e.g. [4, 28]) at a time to increase the joint likelihood and updating the occlusion map accordingly. However, the verification of each hypothesis was performed only once. Thus the hypothesis can be rejected or accepted without considering of the global configuration of the available hypotheses. Recently, Huang and Nevatia [31] proposed a so-called "dynamic" search which allows a hypothesis to be added or removed more than one time. In [32], 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 relationships between the detected parts. However, these rules do not consider the poses and views of the detected parts.

3. A Part-based Human Detector

Since occluded humans are not fully visible, part-based human detectors are suitable for the task of detecting humans in occlusion. In this paper, the human detector proposed by the authors in [33] is used. The detector employs a shape-appearance based human descriptor and SVM classifier for describing and classifying human objects respectively. Readers are referred to [33] for more details.

The shape-appearance based human descriptor can be briefly described as follows. A set of contour templates are used to model the shape of a human body. In order to cope with the occlusion and the variation of human postures and viewpoints due to the articulation of human body, part-based templates are employed. In particular, a template model $\mathcal{M} = \{P_1, P_2, ..., P_N\}$ is a collection of N sets of part templates in which each template $T \in P_i$ represents the shape of part i observed at a certain posture and viewpoint. Figure 3 shows the part templates used in this paper with $N = 4$ and $|P_{top}| = 5$, $|P_{bottom}| = 8$, $|P_{left}| = |P_{right}| = 6$ where $|P_i|$ denotes the cardinality of the set P_i. As shown in Fig. 3, the number of templates to be matched is $5 + 8 + 6 + 6 = 25$ (templates) to cover up to $5 \times 8 \times 6 \times 6 = 1440$ different postures. Compared with full body detection approach, the advantage of part-based detection is that the matching is performed on a small set of templates but covers a variety of human postures.

For each human object image, e.g. a detection window W, a set of best matching part templates

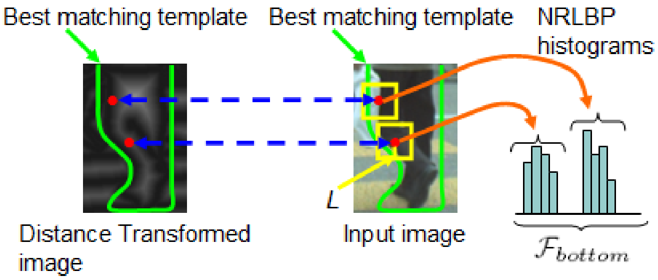


Figure 4: Forming the feature vector.

$\{T_i^*\}, i = 1, 2, ..., N$ is determined as,

$$T_i^* = \arg\min_{T \in P_i} D(T, W) \tag{1}$$

where $D(T, W)$ is the Chamfer distance between a template $T \in P_i$ and the edge map, e.g. Canny's edge map, of window W. In this paper, the generalised distance transform proposed in [10] is used for matching templates.

For each best matching part template T_i^*, a set of edge points E_i on the edge map of W and closest to T_i^* is sampled. This set defines the locations at which the non-redundant local binary patterns (NRLBP) proposed in [34] are extracted. The NRLBP is a variant of the popular LBP [22] and defined as follows. Given a pixel c and S is the number of neighbouring pixels whose the spatial distance to c does not exceed R, we define,

$$NRLBP_{S,R}(c) = \min\left\{LBP_{S,R}(c), 2^S - 1 - LBP_{S,R}(c)\right\} \tag{2}$$

where $LBP_{S,R}(c)$ is the LBP code of c and calculated as,

$$LBP_{S,R}(c) = \sum_{p=0}^{S-1} f(g_p - g_c)2^p \tag{3}$$

where g_p and g_c are the intensities of p and c; $f(x) = 1$ if $x \geq 0$, and $f(x) = 0$, otherwise.

As shown in [34], the NRLBP is more discriminative and offers lower dimension than the original LBP. In addition, the NRLBP is robust and adaptive to changes of the background and foreground.

For each edge point $e \in E_i$, the NRLBP histogram h_e of a $(2L + 1) \times (2L + 1)$-local region centered at e is computed on the detection window W. The part descriptor $\mathcal{F}_i$ is formed as follow,

$$\mathcal{F}_i = \bigoplus_{e \in E_i} h_e \tag{4}$$

where $\bigoplus$ is a concatenating operator. Figure 4 illustrates the formation of the part descriptor, e.g. the bottom part of a human object.

The overall part-based human descriptor $\mathcal{F}$ can be simply constructed by concatenating the descriptors

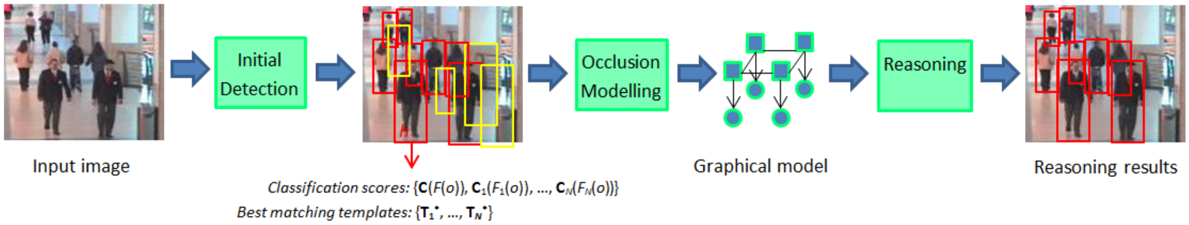

Input image
Initial Detection
h
Classification scores: {C(F(o)), C₁(F₁(o)), ..., C_N(F_N(o))}
Best matching templates: {T₁*, ..., T_N*}
Occlusion Modelling
Graphical model
Reasoning
Reasoning results

Figure 5: Occlusion reasoning framework. Each human hypothesis generated by initial detection is represented by a set of full and part detection scores and a set of best matching templates.

of all parts as,

$$\mathcal{F} = \bigoplus_{i \in \{1,2,\ldots,N\}} \mathcal{F}_i \tag{5}$$

$\mathcal{F}$ and $\mathcal{F}_i, i \in \{1, ..., N\}$ of positive and negative samples are created and used to train a set of classifiers (e.g. SVMs) $\mathbf{C}$ and $\mathbf{C}_i, i \in \{1, ..., N\}$ for classifying the full human body and body parts respectively.

4. Occlusion Reasoning Formulation

Suppose that an initial set of hypotheses about the presence of humans (bounding boxes), $H = \{h_1, h_2, ..., h_X\}$ are detected from an image I using a part-based human detector (e.g. the one presented in the above section). Since at the initial stage, there is no information on whether or which parts of the human body are occluded, the initial set of hypotheses is created by selecting hypotheses h_i that satisfy $\mathbf{C}(\mathcal{F}^{l(h_i)}) \geq \phi$. Here, $\mathcal{F}^{l(h_i)}$ is the part-based feature vector describing a candidate object h_i at its location $l(h_i)$ defined in (5); $\mathbf{C}(\mathcal{F}^{l(h_i)})$ denotes the detection score (i.e. the classification score); and ϕ is a detection threshold which represents the trade-off between true detections and false alarms. The threshold, ϕ, is set for a conservative detection such that true positives are not missed and false positives may be included. The inter-object occlusion of hypotheses is then represented as a graphical model G on which the reasoning is formulated. Figure 5 illustrates the reasoning process.

The graphical model $G(V, E)$ where V and E are the vertices and edges can be created as follows. For each hypothesis $h_k, k \in \{1, ..., X\}$, a binary value is assigned to indicate that the hypothesis h_k is a false positive (value of 0) or true positive (value of 1). Let o_k denote the corresponding image region of a hypothesis h_k. Given H, the corresponding image data $O = \{o_1, o_2, ..., o_X\}$ can be obtained. As the values of $h_k, k \in \{1, ..., X\}$ are to be determined, we treat them as hidden nodes (i.e. state variables) while the image data $o_k, k \in \{1, ..., X\}$ are considered as observable nodes. Finally, we define $V = \{H, O\}$.

For the edges E, there are two types: observation-state edge and state-state edge. The observation-state edges connecting h_k and o_k represent the observation likelihood $p(o_k|h_k)$. For overlapping hypotheses, there may exist an inter-object occlusion relationship, thus a link between those hidden nodes, i.e. state-state edge, is added. It has been observed that, because of the perspective rule, if h_k is occluded by h_j then the image region of h_j is larger than that of h_k and foot position of h_j is higher than that of h_k (with image coordinate shown in Figure 2). This observation is reasonable and valid in most surveillance systems, e.g. [4, 5, 27, 30]. In our model, this property is exploited in defining state-state edges. In particular, if h_k is occluded by h_j (determined by the foot positions and/or image areas), there is an

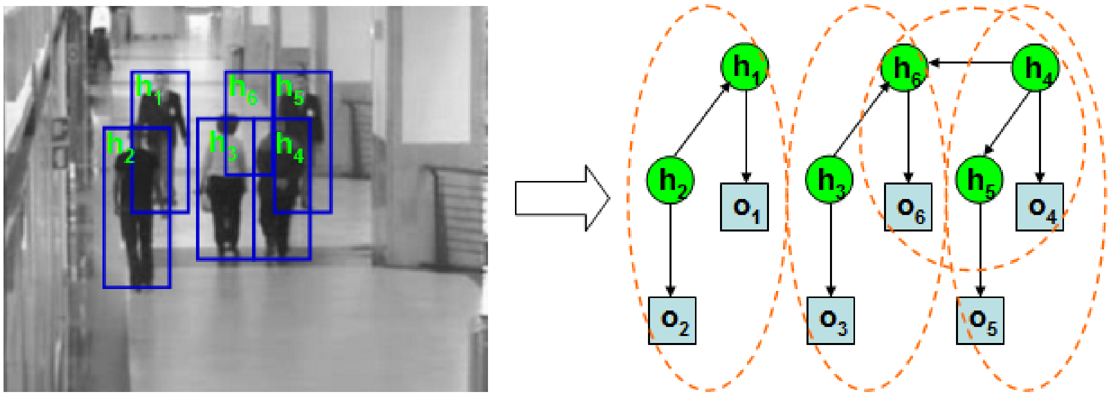


Figure 6: Left: initial detection hypotheses, Right: the graphical model in which ellipses represent local groups.

edge from h_j to h_k with a probability $p(h_k|h_j)$. This design implies that the presence of h_j affects the detection of h_k, but not vice versa, since h_j occupies the view space of h_k from the camera. Figure 6 shows an example of the graphical model G. As can be seen, G can be considered as a Bayesian network in which h_k and o_k are state and observed variables respectively.

The initial set of hypotheses, H, may contain false alarms (i.e. incorrect assignment of values to h_k). This is because the initial detection is performed based on recognising individual image windows independently without considering the geometrical layout of humans in the scene. Note that refining this set is considered as making inference on appropriate values of hypotheses h_k in estimating the marginal probability of the observed data:

$$\log p(O) = \log \int_H p(O|H)p(H)dH \tag{6}$$

where $p(H)$ is the prior and $p(O|H)$ is the likelihood of obtaining the observed data O given states H.

Since each h_k takes a binary value, a brute force estimation of (6) would require $O(2^X)$ operations. In the next section, an algorithm is proposed to effectively estimate $\log p(O)$ using the variational mean field method.

5. Reasoning Using Variational Mean Field Method

5.1. Variational Approach

Variational methods have been commonly used for approximate inference and estimation. Basically, in variational approach, the estimation problem is casted as optimisation problem and an approximate solution is to be found [35]. This is advantageous especially when the exact solution is not feasible or practical to obtain. In our problem, instead of estimating $\log p(O)$ using (6), we can approximate it by optimising an objective function $J(Q)$ of a variational distribution Q as follows,

$$J(Q) = \log p(O) - KL(Q(H)||p(H|O)) \tag{7}$$

where KL is the Kullback-Leibler divergence of two distributions [36] defined as,

$$KL(Q(H)||p(H|O)) = \int_H Q(H) \log \frac{Q(H)}{p(H|O)} dH \tag{8}$$

Substituting (8) into (7), we obtain,

$$\begin{aligned} J(Q) &= \log p(O) - \int_H Q(H) \log Q(H) dH + \int Q(H) \log p(H|O) dH \\ &= \log p(O) - \int_H Q(H) \log Q(H) dH + \int_H Q(H) \log \frac{p(H,O)}{p(O)} dH \\ &= - \int_H Q(H) \log Q(H) dH + \int_H Q(H) \log p(H,O) dH \\ &= \mathcal{H}(Q) + E_Q\{\log p(H,O)\} \end{aligned} \tag{9}$$

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 respect to Q will give us $J(Q^*)$ as an approximation of $\log p(O)$. Note that Q^* will also be an approximate of the posterior $p(H|O)$. In addition, an approximation of $\log p(O)$ corresponds to finding an appropriate variational distribution $Q(H)$. In this paper, the simplest variational distribution that all hidden variables are assumed to be independent of each other is adopted. In particular, we assume,

$$Q(H) = \prod_{k=1}^{X} Q_k(h_k) \tag{10}$$

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

$$\mathcal{H}(Q) = \sum_{k=1}^{X} \mathcal{H}(Q_k) \tag{11}$$

where $\mathcal{H}(Q_k)$ is the entropy of Q_k.

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

$$J(Q) = \text{const.} + \mathcal{H}(Q_k) + \int_{h_k} Q_k(h_k) E_Q\{\log p(H,O)|h_k\} \tag{12}$$

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

As presented in [35], maximising $J(Q)$ can be obtained by computing Gibbs distributions of $Q_k(h_k)$:

$$Q_k(h_k) \leftarrow \frac{1}{Z_k} e^{E_Q\{\log p(H,O)|h_k\}} \tag{13}$$

where Z_k is the normalisation factor computed as,

$$Z_k = \int_{h_k} e^{E_Q\{\log p(H,O)|h_k\}} \tag{14}$$

Update equations (13) and (14) will be invoked iteratively to increase the objective function $J(Q)$. As can be seen, the computation of (13) and (14) requires an estimation of $E_Q\{\log p(H,O)|h_k\}$ which is dependent on the configuration of the graphical model. This will be presented in the next section.

5.2. Computation

It is observed that the update of $E_Q\{\log p(H,O)|h_k\}$ depends only on h_k and hypotheses occluded by h_k. In essence, the presence of a node h_k affects only the likelihood $p(o_k|h_k)$ explaining how likely we have h_k given observation data o_k and likelihoods of nodes occluded by h_k. Thus, $E_Q\{\cdot|h_k\}$ can be factorised over a set of *local groups* containing nodes related to h_k where the update can be performed locally. In particular, let $\mathcal{N}(h_k)$ be the set of neighbouring hidden nodes of h_k, i.e. hidden nodes which are directly connected to h_k. For each node $h_j \in \mathcal{N}(h_k)$, a *local group* c representing the dependency of h_j on h_k is defined as $c_{k,j} = (h_k, o_k, h_j, o_j)$. Figure 6 shows an example of groups. The update can be performed simply as,

$$E_Q\{\log p(H,O)|h_k\} \leftarrow \sum_{h_j \in \mathcal{N}(h_k)} \int_{h_j} Q_j(h_j) \log \psi(c_{k,j}) \tag{15}$$

where $\psi(c_{k,j})$ is the potential function of the *group* $c_{k,j}$. It can be computed as in a conventional Bayesian network:

$$\psi(c_{k,j}) \equiv p(h_k, o_k, h_j, o_j) = p(o_k|h_k)p(h_j|h_k)p(o_j|h_j)p(h_k) \tag{16}$$

where $p(o_k|h_k)$ represents the likelihood of the human hypothesis h_k given the observation o_k. $p(h_k)$ is the prior of the presence of h_k and $p(h_j|h_k)$ indicates a state transition in a Bayesian network.

Since H is revised iteratively, (15) needs to be updated accordingly with current settings of H. For example, at each time and based on a particular setting of H, for each hypothesis h_k, $k \in \{1, ..., X\}$, visible parts are determined using (21) and the likelihood $p(o_k|h_k)$ in (16) can be re-evaluated. To compute $p(o_k|h_k)$, the detection scores of visible parts can be used. However, since the parts of a human object are detected independently, they may not represent any regular configuration of a human body. Fortunately, with the shape-appearance human detector, it is possible and worthwhile to validate the combination of parts using the best matching templates $\{T_i^*\}$ computed in (1). In particular, we define the likelihood in $p(o_k|h_k)$ as,

$$p(o_k|h_k) = \begin{cases} \phi, & \text{if } h_k = 0 \\ p(\{T_i^*\}, \{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k), & \text{if } h_k = 1 \end{cases} \tag{17}$$

where ϕ is the detection threshold used to obtain the initial set of hypotheses; $\{T_i^*\}$ and $\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}$ are the sets of best matching part templates and corresponding part detection scores (e.g. the classification scores of SVMs in our experiments) of visible parts of the observed image data o_k. Those sets can be used to interpret the observation o_k.

Assuming that $\{T_i^*\}$ and $\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}$ are statistically independent variables given h_k, (17) can be rewritten as,

$$p(o_k|h_k) = \begin{cases} \phi, & \text{if } h_k = 0 \\ p(\{T_i^*\}|h_k)p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k), & \text{if } h_k = 1 \end{cases} \tag{18}$$

To validate the combination of detected parts, $p(\{T_i^*\}|h_k)$ is considered as the co-occurrence of part templates that represents a valid human posture (as $h_k = 1$). In this paper, we model $p(\{T_i^*\}|h_k)$ using a sigmoid function. This is because, the sigmoid function is able to reflect the fact that the higher co-occurrence of part templates is (in the training samples), the more confident those parts form a regular human posture. We define,

$$p(\{T_i^*\}|h_k = 1) = \frac{1}{1 + e^{-mf(\{T_i^*\})}} \tag{19}$$

where m is an empirical parameter and $f(\{T_i^*\})$ is the frequency with which visible parts $\{T_i^*\}$ co-occur. Note that $p(\{T_i^*\}|h_k = 1)$ can be computed off-line and retrieved from a look-up table manner.

The term $p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k = 1)$ in (18) can be calculated as,

$$p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k = 1) \propto \sum_{i=1}^{N} v_i \mathbf{C}_i(\mathcal{F}_i(o_k)) \tag{20}$$

where v_i is a binary parameter indicating whether the part i is occluded or not. More precisely, we define,

$$v_i = \begin{cases} 1, & \text{if } occ(i) < \delta \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

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, $p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k = 1)$ defined in (20) is the sum of the part detection scores of visible parts of a hypothesis h_k. To make the likelihood $p(o_k|h_k)$ invariant to the number of visible parts, $p(\{T_i^*\}|h_k = 1)$ is used to compensate $p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k = 1)$ in (18) when occlusion occurs. For instance, with the part model $\mathcal{M}$ used in [33] (i.e. a human body is decomposed into $N = 4$ parts including top, bottom, left, and right) and assuming that only the bottom part is occluded, $p(\{\mathbf{C}_i(\mathcal{F}_i(o_k))\}|h_k = 1), i \in \{top, left, right\}$ would decrease while $p(T_{top}^*, T_{left}^*, T_{right}^*|h_k = 1) > p(T_{top}^*, T_{bottom}^*, T_{left}^*, T_{right}^*|h_k = 1)$. In implementation, the prior of the configurations $p(\{T_i^*\}|h_k = 1)$ can be pre-computed and accessed from a look-up table; thus there is no computational overhead associated with its usage.

Note that when other human detectors not using template matching, e.g. [7] are employed, the term $p(\{T_i^*\}|h_k = 1)$ simply reflects the probability of the presence of the parts and hence can be computed as $p(\{T_i^*\}|h_k = 1) = \frac{1}{\sum_{i=1}^N v_i}$. The likelihood $p(o_k|h_k)$ then becomes $\frac{\sum_i^N v_i \mathbf{C}_i(\mathcal{F}_i(o_k))}{\sum_{i=1}^N v_i}$, i.e. the average of the part detection scores of visible parts as used in [24].

To compute (15), we assume that $p(h_k = 0) = 1 - p(h_k = 1) = \rho$ and $p(h_j|h_k) = \varrho$ for all $h_k, h_j \in \{0, 1\}$. In addition, if h_k does not occlude any other hypotheses (e.g. h_1 in Figure 6), $\psi(c)$ will be simplified to $p(o_k|h_k)$ as the presence of h_k does not affect any other hypotheses. This means that $E_Q\{\log p(H, O)|h_k\}$ depends only on the likelihood of h_k to the observation o_k. We initialised $Q_k(h_k = 1) = 1 - Q_k(h_k = 0) = \nu$. Finally, if $Q_k(h_k = 1) \geq Q_k(h_k = 0)$, h_k is set to 1, (i.e. true detection) and $p(o_k|h_k)$ is re-evaluated using (18) with current setting of h_k. When the optimal Q^* is found, the corresponding subset of hypotheses $h_k = 1$ is determined. This subset provides the final detection results. The proposed reasoning algorithm is summarised in Algorithm 1.

Unlike greedy-based occlusion reasoning (e.g. [4, 28, 5, 29]), our method tests each hypothesis and its occlusion status is verified (by adding/removing) more than once. The presence (state) of each hypothesis is determined by its likelihood and the likelihoods of its neighbours to maximise the objective function. Such a reasoning method avoids rejecting hypotheses too early as in the greedy approach where each hypothesis has no chance to be reconsidered once it has been rejected. In addition, compared with some

Algorithm 1 Reasoning Algorithm

```
$stop = FALSE$
$J(Q^*) = 0$
while $stop == FALSE$ do
  $J(Q) \leftarrow 0$
  $stop \leftarrow TRUE$
  for $k = 1$ to $X$ do
    $\mathcal{H}(Q) \leftarrow \mathcal{H}(Q) - \mathcal{H}(Q_k)$
    Update $Q_k(h_k)$ and $E_Q\{\cdot|h_k\}$
    $\mathcal{H}(Q) \leftarrow \mathcal{H}(Q) + \mathcal{H}(Q_k)$
    $J(Q) \leftarrow \mathcal{H}(Q) + \int_{h_k} Q_k(h_k) E_Q\{\cdot|h_k\}$
    if $J(Q) > J(Q^*)$ then
      $stop \leftarrow FALSE$
      $J(Q^*) \leftarrow J(Q)$
      if $Q_k(h_k = 0) > Q_k(h_k = 1)$ then
        $h_k \leftarrow 0$
      else
        $h_k \leftarrow 1$
      end if
    end if
  end for
end while
```

window-based occlusion reasoning methods, e.g. [23, 24], our proposed method offers a general framework where full and part detectors can be implemented using different types of features, object representations, and classifiers. In contrast, the method in [23] depends on a grid-based object representation and a linear SVM and in [24] motion and depth information (stereo images) are required for reasoning.

6. Experimental Results

6.1. Experimental Setup

There are a number of parameters used in the human detector and occlusion reasoning algorithm. Values of the parameters of the part-based human descriptor were set similarly to [33]. In particular, for a 96-pixel tall human, the window size L of local image regions centered at contour points was set to 7. The number of neighbouring pixels S and the radius R in (2) were 8 and 1 respectively. Details of parameter setting of the human detector was presented in [33]. For parameters related to occlusion reasoning, without any prior knowledge about occlusion, we set δ defined in (21) to 0.5, $\rho = p(h_k = 0) = 1 - p(h_k = 1) = 0.5$, $\varrho = p(h_j|h_k) = 0.5$, and $\nu = Q_k(h_k = 1) = 1 - Q_k(h_k = 0) = 0.5$. In addition, ϕ is varied to represent the trade-off between true detections and false alarms. We have tried the parameters with different values but the performance was slightly different while the above setting gave the best performance.

The part-based and part detectors were trained independently. The training of each detector involves two steps: initial training and bootstrapping. For the initial training, the training set is from the INRIA dataset [37] and consisted of 2416 positive samples and 12180 negative samples (created by selecting randomly 10 samples per negative image). In the bootstrapping, the negative images were exhaustedly searched to find the 2300 hard-to-detect negative samples whose positive probability is higher than a predefined threshold (set to 0.2 in our experiments). The hard negative samples together with the original positive and negative samples were used to train each detector once again.

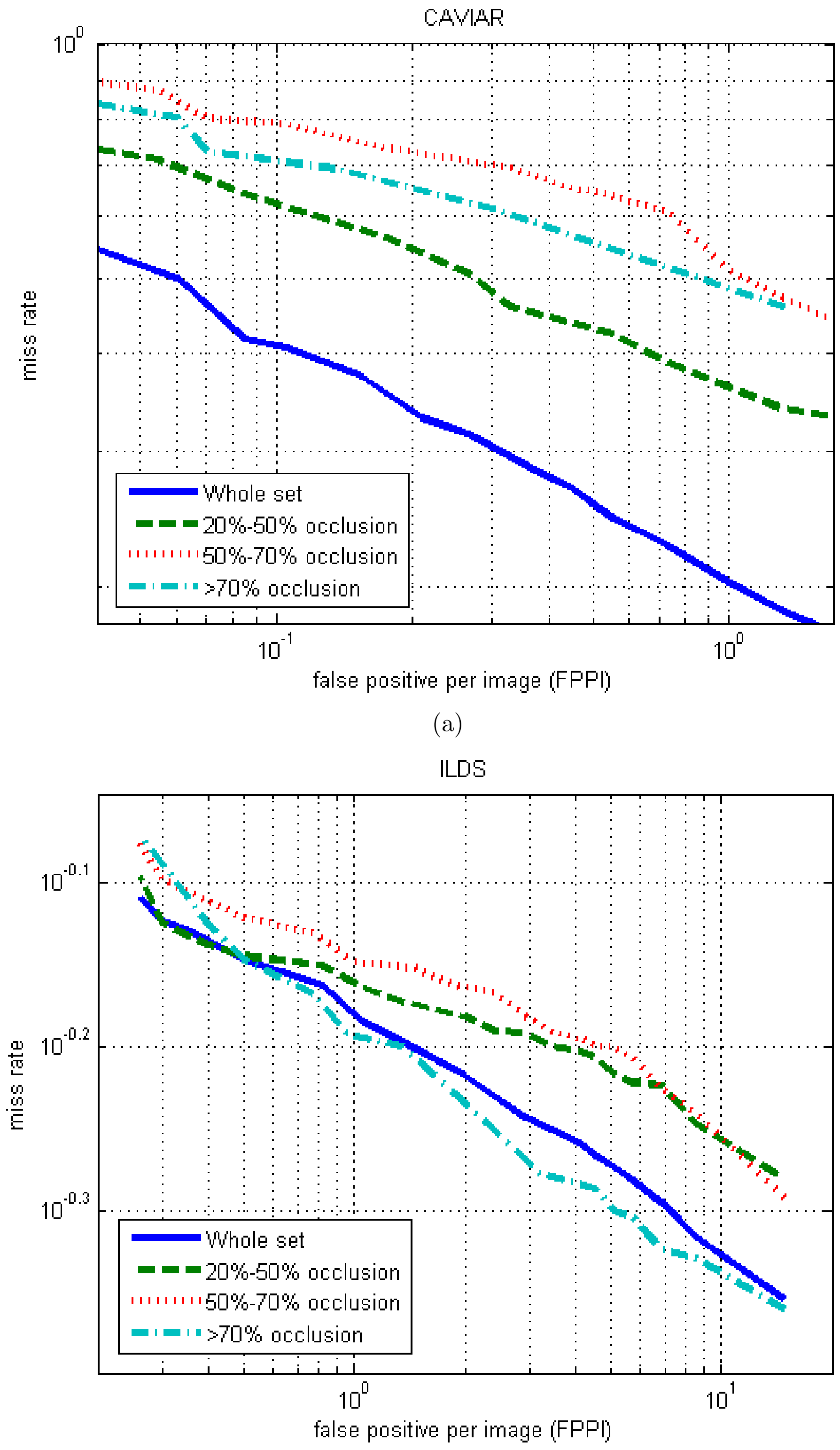
6.2. Performance Evaluation

The proposed algorithm was evaluated on three datasets A, B, and the test set of the INRIA dataset. The set A was created by selecting 200 images (800th-1000th frame) of 384×288-pixels with 1614 annotated humans from the *OneStopMoveEnter1cor* sequence of the CAVIAR dataset [38]. The set B contains 301 images of 720×576-pixels extracted from the *Hard* sequence of the iLIDS dataset [39]. On this set, we labelled 3314 humans. Compared with the set A, set B is more challenging due to the high level and the variation of occlusion. The robustness of the proposed method was also verified in detecting non-occluded humans. The test set of the INRIA dataset [37] was used for this case.

On the datasets A and B, the evaluation was conducted at various levels of occlusion and the whole dataset (i.e. all levels of occlusion). For occluded humans, we evaluated the reasoning method based on different levels of occlusion including: 20%-50%, 50%-70%, and more than 70%. The occlusion level of a human object was computed as the ratio of the occluded area and the area of the tightly enclosing ellipse of the human object. The detection error trade-off (DET) measure computed based on false positive per image (FPPI) versus miss rate was used as the evaluation measure [2]. In general, precision-recall is often used for object detection. However, for evaluating the detection performance on only occluded humans given the annotation of non-occluded and occluded humans, the DET is more appropriate than the precision-recall measure. This is because the precision is computed relatively to the number of false alarms and, on the detection of occluded humans, only occluded humans are considered and the number of positives may be much smaller than the number of false alarms. Figure 7 shows the detection performance on the whole set and occluded humans of both test sets A and B. It can be seen that on both of the test sets, the reasoning method obtains poor performance when the occlusion reaches 50%-70%. However, on the set B, the reasoning method achieves the best detection performance at more than 70% of occlusion, i.e. the detection performance is even better than that estimated on the whole set. An illustration of the reasoning process with interim results on the dataset A is presented in Figure 8. As can be seen from Figure 8, at each step of the reasoning, false alarms are removed and true detections are recovered. Figure 9 shows some detection results.

On the INRIA dataset, 288 full images were used instead of cropped positive and negative samples. This is because the purpose of the experiment is to evaluate the robustness of the occlusion reasoning algorithm in improving the detection accuracy. In addition, occlusion inference is performed based on the spatial layout of detection hypotheses in the scene. Figure 10 shows the detection performance on the INRIA dataset with and without using the reasoning algorithm. Some detection results are presented in Figure 12. Interestingly, through experiments we have found that although most of humans in this set are not occluded, the reasoning algorithm did not introduce any adverse effect and it somehow improved the detection performance. This is probably because, as shown in Figure 12, in cases where a false alarm overlaps with a true detection the reasoning process could infer possible occluded parts to invoke proper part detectors to verify the false alarms. In addition, the reasoning method makes the justification of a hypothesis as false alarm or true detection based on not only the hypothesis itself but also on its contribution to a global configuration of spatially related hypotheses.

As the reasoning algorithm is an iterative process, the efficiency of the proposed method needs to be evaluated. Recall that the occlusion inference is performed iteratively to maximise the objective function $J(Q)$ and at each iteration all hypotheses are verified and the 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. Those



(a)

(b)

Figure 7: Detection performance on the dataset A (a) and B (b).

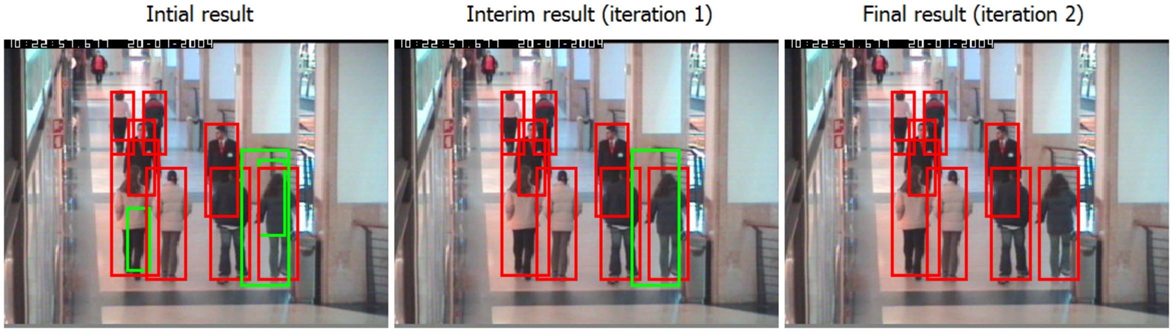


Figure 8: An illustration of occlusion reasoning in which green rectangles represent false alarms.

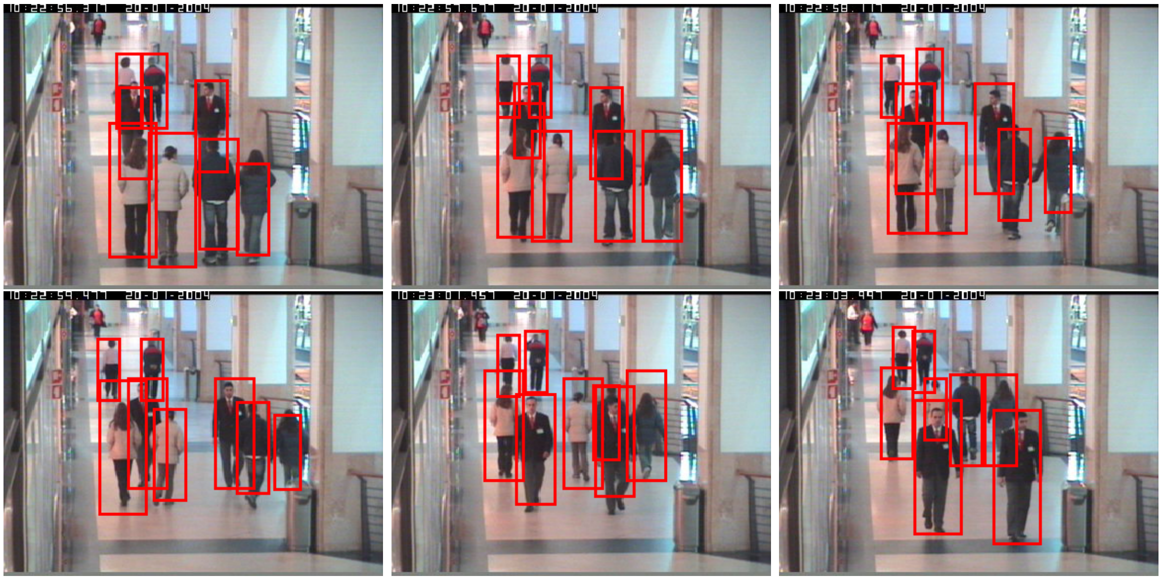


Figure 9: Some results of human detection in occlusion.

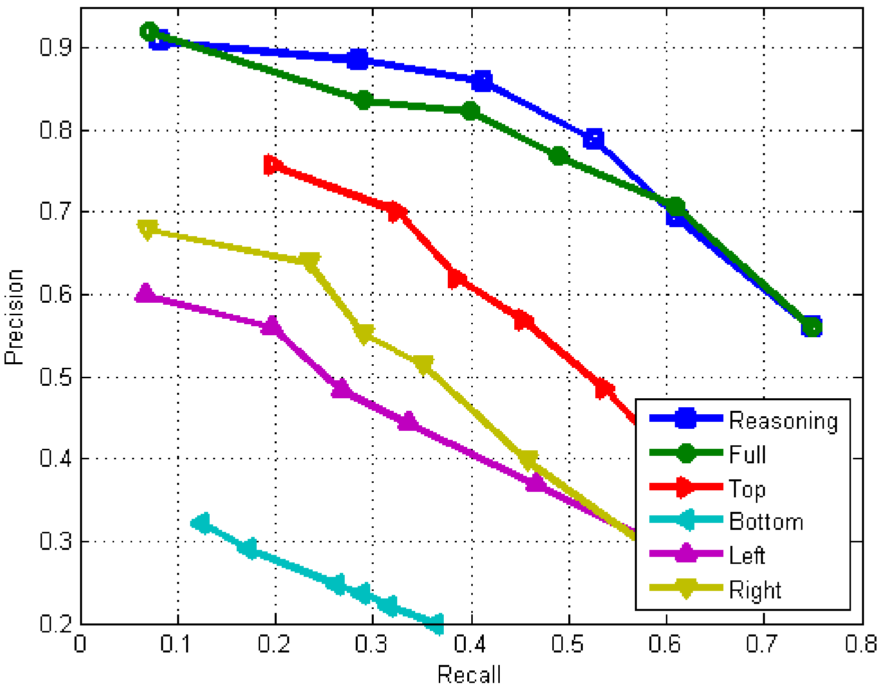


Figure 10: Detection performance of the full detector, part detectors, and reasoning method on the INRIA dataset.

numbers depend on the number of hypotheses and the frequency of occlusions. On the average, through experiments, we have found that the number of iterations, e.g. on over 200 images of the set A, is about 2.1 and each 384×288 frame can be processed in approximately 0.25 seconds for the occlusion reasoning.

6.3. Comparison

To verify the robustness of the reasoning algorithm, we compared the human detector with and without using the reasoning method. The comparison is shown in Figure 11. The log-average miss rate (LAMR) proposed in [40] was used as the evaluation measure. The LAMR of a method is computed by averaging the miss rates at different FPPI rates in the range from 10^{-1} to 10^{0}. A low value of the LAMR indicates better performing detection method. Through experiments, we have found that, the reasoning method could improve the detection performance. For example, the LAMR of the detector without reasoning was about 0.46 while it was 0.32 by using the reasoning method (i.e. a reduction by approximately 14%).

The proposed reasoning algorithm was also compared with other algorithms. In particular, the reasoning methods proposed by Wu and Nevatia in [4, 28], by Lin et al. in [5, 29] (and then used by Beleznai and Bischof in [30]) and by Huang and Nevatia [31] were selected for comparison. In [4, 28], all detection responses were hypothesized initially, and the inferencing process was conducted by removing false candidates. On the other hand, Lin et al. [5, 29] started with an empty set of hypotheses and extended this set by adding true candidates. Essentially, in this method, deciding whether a hypothesis is true positive or false alarm is based only on the image likelihood of the hypothesis itself. In [31], each hypothesis can be added and removed more than one time and the search process, called ”dynamic

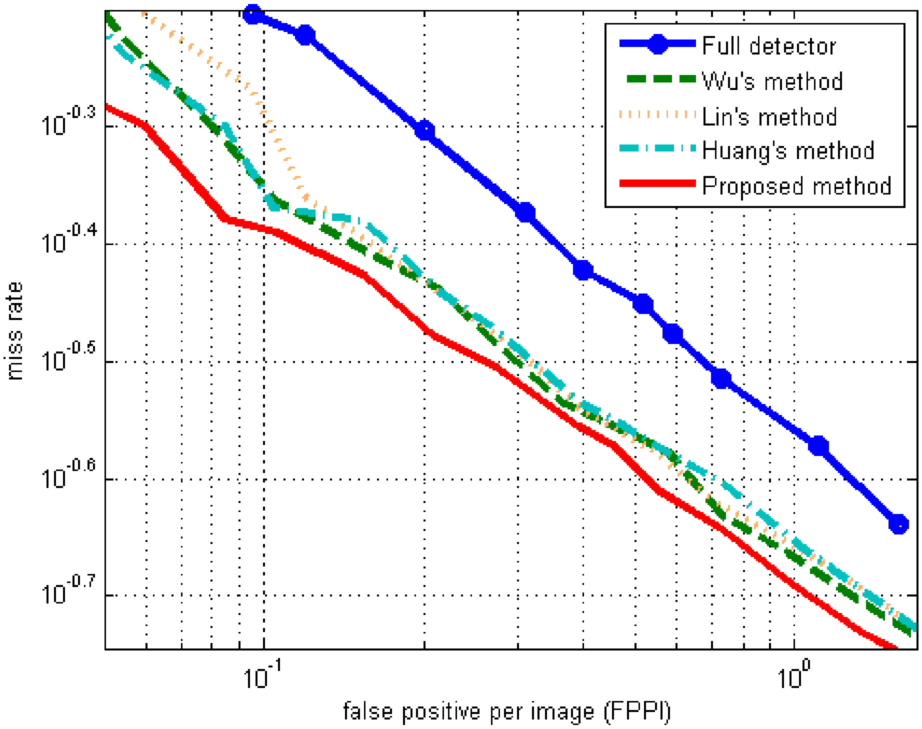


Figure 11: Comparison with full detector and Lin's method [29].

search" to compare with "static search" in [4, 28], is terminated when there is no any improvement gained by adding/removing hypotheses. To obtain a fair comparison, we used the same human detectors for all competitive methods. In addition, for those methods, the joint likelihood of hypotheses and experimental parameters were computed similarly to [5]. Readers are referred to [5] for more details[1].

Figure 11 shows the comparison of all methods. As can be seen from this figure, in general, there was a slight difference in the detection performance between the works in [4, 28] and in [5, 29] in the range of $[10^-1, 10^0]$ of FPPI. The same situation can also be found for the method of Huang and Nevatia [31] though it was claimed in their work that "dynamic search" outperformed "static search" in [4, 28]. However, experimental results have shown that our method obtained better detection performance compared with all of those methods. In particular, the LAMR of our method was 0.32 while it was 0.35 for Wu's method, 0.39 for Lin's method, and 0.36 for Huang's method.

7. Discussion and Conclusion

~~An issue of the proposed reasoning algorithm is the accuracy of the approximate solution.~~ Theoretically, the accuracy of variational approximation can be considered as the difference $\log p(O) - J(Q)$ or the tightness of the variational marginals $\{Q_k(h_k)\}$ on the true posterior marginals $p(h_k|O)$. However, as shown in [35], these two criteria do not always agree with each other and this depends on the structure of

[1] For the method in [4, 28], since parts of a human hypothesis were detected simultaneously, matching detection responses to hypotheses was not computed and thus "false negatives" were not used in calculating the joint likelihood.

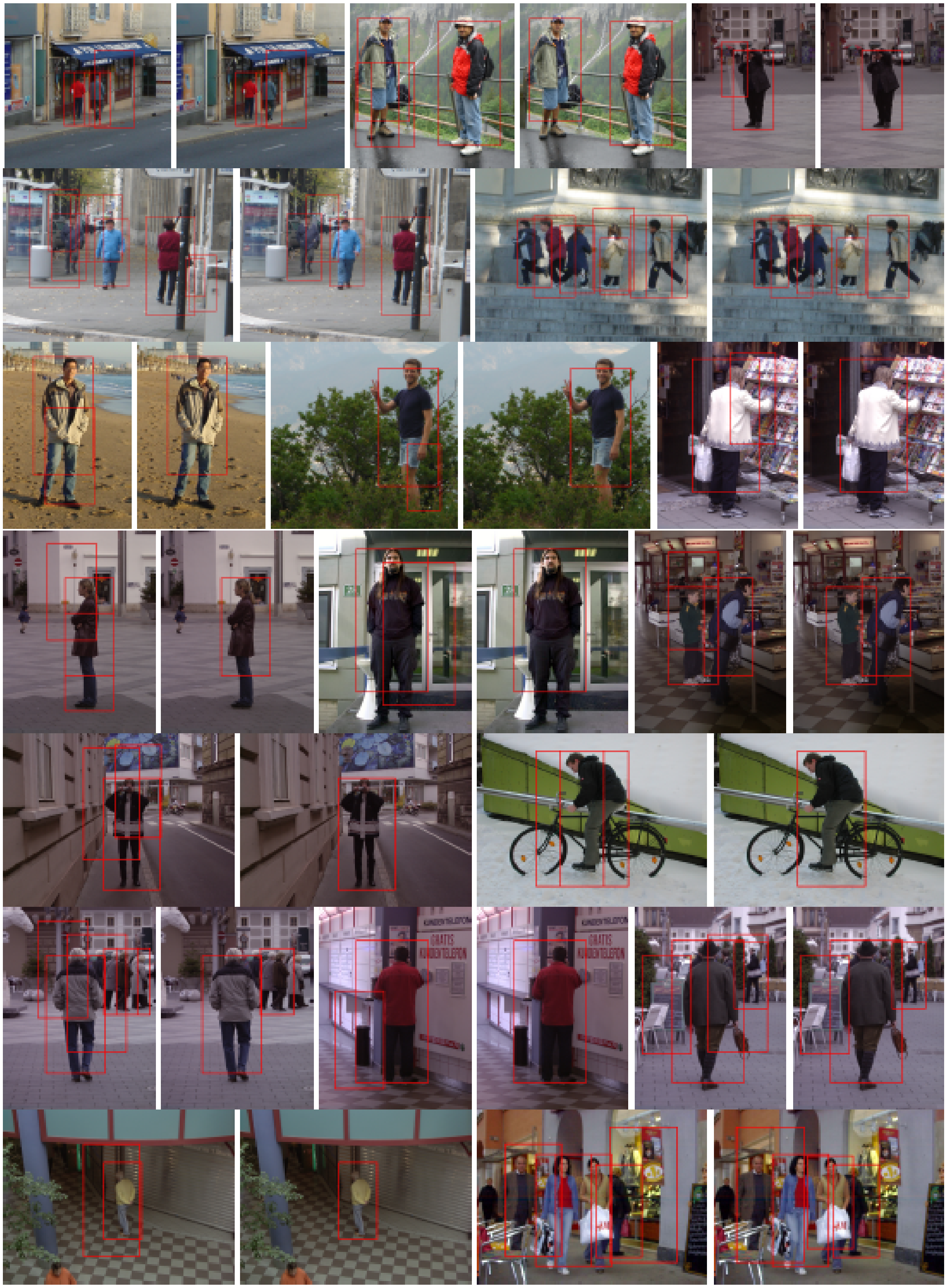


Figure 12: Illustration of occlusion reasoning on the INRIA dataset. For each pair of images, the left image shows the initial detection and the right image shows the final detection result after reasoning.

the graphical model. In general, graphical models with weak dependencies between nodes are expected to have good approximate solution compared with models with strong dependencies between nodes. In the problem of inter-object occlusion reasoning, we have found that, the existence of a hypothesis affects only its nearby hypotheses. Furthermore, we also have observed that, in practice, a human object, to be identified, is often occluded by few (e.g. no more than three) other human objects. Thus, we could expect that a good approximation can be obtained by the variational mean field method. This explains the success of the proposed method in inter-object occlusion reasoning. There also exit some other inference algorithms for graphical models, e.g. loopy belief propagation (LBP) used in Conditional Random Field. Those methods will be investigated in our future work.

Recently, a number of part-based human detectors, in which locations of parts are also indicated, have been developed, e.g. [41]. The graphical model and variational mean field method could be applied to model not only the interaction between human objects but also the spatial relationship between parts of a human object. The advantage of this approach is its extension in solving self-occlusion through inferring the human poses.

In all, this paper proposes an inter-object occlusion reasoning algorithm based on variational mean field for detecting multiple partially occluded humans. The proposed algorithm can accommodate various human detectors using different features, object representations and classifiers. The inter-object occlusion is modelled as a graph and the occlusion reasoning is formulated as estimation of a marginal probability of the observed data in a Bayesian network. The reasoning algorithm was evaluated on the different datasets and experimental results show the robustness and efficiency of the proposed method not only in detecting humans under severe occlusion but also in the cases where there is no occlusion.

References

[1] M. Enzweiler, D. M. Gavrila, Monocular pedestrian detection: Survey and experiments, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12) (2009) 2179–2195.

[2] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 304–311.

[3] B. Wu, R. Nevatia, Cluster boosted tree classifier for multi-view, multi-pose object detection, in: Proc International Conference on Computer Vision, 2007, pp. 1–8.

[4] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Proc International Conference on Computer Vision, 2005, pp. 90–97.

[5] Z. Lin, L. S. Davis, D. Doermann, D. DeMenthon, Hierarchical part-template matching for human detection and segmentation, in: Proc International Conference on Computer Vision, 2007, pp. 1–8.

[6] D. M. Gavrila, V. Philomin, Real-time object detection for smart vehicles, in: Proc IEEE International on Computer Vision, Vol. 1, 1999, pp. 87–93.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 886–893.

[8] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (4) (2001) 349–361.

[9] D. M. Gavrila, A Bayesian, exemplar-based approach to hierarchical shape matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (8) (2007) 1–14.

[10] D. T. Nguyen, W. Li, P. Ogunbona, An improved template matching method for object detection, in: Proc Asian Conference on Computer Vision, Vol. 3, 2009, pp. 193–202.

[11] J. Shotton, A. Blake, R. Cipolla, Contour-based learning for object detection, in: Proc International Conference on Computer Vision, Vol. 1, 2005, pp. 503–510.

[12] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of adjacent contour segments for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (1) (2008) 36–51.

[13] Q. Zhu, S. Avidan, M. C. Yeh, K. T. Cheng, Fast human detection using a cascade of histograms of oriented gradients, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1491–1498.

[14] P. Sabzmeydani, G. Mori, Detecting pedestrians by learning shapelet features, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[15] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on riemannian manifolds, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[16] W. Gao, H. Ai, S. Lao, Adaptive contour features in oriented granular space for human detection and segmentation, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 1786–1793.

[17] P. Viola, M. J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. 511–518.

[18] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 878–885.

[19] Y. Mu, S. Yan, Y. Liu, T. Huang, B. Zhou, Discriminative local binary patterns for human detection in personal album, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[20] P. Ott, M. Everingham, Implicit color segmentation features for pedestrian and object detection, in: Proc International Conference on Computer Vision, 2009, pp. 723–730.

[21] P. Viola, M. J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63 (2) (2005) 153–161.

[22] T. Ojala, M. Pietikăinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29 (1) (1996) 51–59.

[23] X. Wang, T. X. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in: Proc International Conference on Computer Vision, 2009, pp. 32–39.

[24] M. Enzweiler, A. Eigenstetter, B. Schiele, D. M. Gavrila, Multi-cue pedestrian classification with partial occlusion handling, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 990–997.

[25] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.

[26] P. Dollár, B. Babenko, S. Belongie, P. Perona, Z. Tu, Multiple component learning for object detection, in: Proc European Conference on Computer Vision, Vol. 2, 2008, pp. 211–224.

[27] T. Zhao, R. Nevatia, B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1198–1211.

[28] B. Wu, R. Nevatia, Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors, International Journal of Computer Vision 75 (2) (2007) 247–266.

[29] Z. Lin, L. S. Davis, Shape-based human detection and segmentation via hierarchical part-template matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (4) (2010) 604–618.

[30] C. Beleznai, H. Bischof, Fast human detection in crowded scenes by contour integration and local shape estimation, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 2246–2253.

[31] C. Huang, R. Nevatia, High performance object detection by collaborative learning of joint ranking of granules features, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 41–48.

[32] V. D. Shet, J. Neumann, V. Ramesh, L. S. Davis, Bilattice-based logical reasoning for human detection, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[33] D. T. Nguyen, P. Ogunbona, W. Li, Human detection with contour-based local motion binary patterns, in: Proc IEEE Conference on Image Processing, 2011, pp. 3609–3612.

[34] D. T. Nguyen, Z. Zong, W. Li, P. Ogunbona, Object detection using non-redundant local binary patterns, in: Proc IEEE Conference on Image Processing, 2010, pp. 4609–4612.

[35] T. S. Jaakkola, Tutorial on variational approximation methods, Tech. rep., MIT Artificial Intelligence Laboratory (2000).

[36] S. Kullback, R. A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 76–86.

[37] http://pascal.inrialpes.fr/data/human/.

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

[39] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html.

[40] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2012) 743–761.

[41] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Proc IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.