An MRF-Poselets Model for Detecting Highly Articulated Humans

Duc Thanh Nguyen[1], Minh-Khoi Tran[2], and Sai-Kit Yeung[3]
Singapore University of Technology and Design
{[1]ducthanh_nguyen,[3]saikit}@sutd.edu.sg, [2]minhkhoi_tran@mymail.sutd.edu.sg

Abstract

Detecting highly articulated objects such as humans is a challenging problem. This paper proposes a novel part-based model built upon poselets, a notion of parts, and Markov Random Field (MRF) for modelling the human body structure under the variation of human poses and viewpoints. The problem of human detection is then formulated as maximum a posteriori (MAP) estimation in the MRF model. Variational mean field method, a robust statistical inference, is adopted to approximate the MAP estimation. The proposed method was evaluated and compared with existing methods on different test sets including H3D and PASCAL VOC 2007-2009. Experimental results have favourbly shown the robustness of the proposed method in comparison to the state-of-the-art.

1. Introduction

Human detection is an active topic in Computer Vision due to its variety of applications such as image/video retrieval, video-based surveillance, driving assistance systems, human-computer interaction in intelligent systems, etc. The problem is also well-known for its challenges including the variation of the human poses, viewpoints, and the occurrence of occlusions.

Literature has shown extensive studies in human detection during a decade [13, 5, 31] and the focuses of the state-of-the-art have been mainly on features and object representation models. Well-known features used in human detection include Haar-like features [34], histogram of oriented gradients (HOG) [11, 16, 7, 6], local binary patterns (LBP) [21], channel features [3, 4, 12], and combination of those features [40, 10]. Recently, deep neural networks [18, 8, 20] have been used for automatic learning of higher-level features. They obtain good results but require specific knowledge and experience to design the network architecture.

Given the features, a detection method represents a human object using global or local models. Global models represent the human object as a whole. The most common representation scheme is the model proposed in [11] in which the human object image (detection window) is represented in a regular grid of blocks. Features, e.g. histograms of oriented gradients (HOG) [11], are then extracted from the blocks and concatenated to form a feature vector describing the whole human object. The success of this approach has been confirmed in detecting human objects with less variation of poses and viewpoints, e.g. pedestrians.

On the other hand, local methods model the human object as a set of parts, e.g. [25, 7, 1, 15, 37, 38, 19, 8, 20]. Compared with the global approach, the local approach is more robust to model the high articulation of the human body. However, the performance of local methods significantly depends on detecting parts and modelling the configuration of the human body based on parts. In heavy deformations, e.g. articulated actions of athletes in sport videos, a part appears differently in different poses and viewpoints; and the spatial distribution of parts also much varies.

1.1. Related Work

In general, the parts in a part-based model can be detected simultaneously [1, 16, 15, 37] or independently [25, 33, 27, 7, 6, 38, 19, 8, 20] in different processes. The most prominent part-based model detecting parts simultaneously is the deformable part model (DPM) proposed by Felzenszwalb et al. [16, 15]. In this model, HOG was used to describe the appearance of the body parts while the placements of the parts were encoded by latent variables of a latent SVM. The parts were detected so as to maximise the appearance score and at the same time minimise the deformation cost. The DPM was then extended in [37] in which more small and rigid parts were used and the appearance consistency between parts was considered. However, the DPM approach requires that the number of parts and part types (e.g. head, torso, legs) must be determined in advance. In fact, due to the variation of poses and viewpoints, the number of parts and part's appearance vary significantly. For example, one arm may not be visible when the human object is in a profile view and the face is not observable in a back-facing pose. In addition, the legs in a running and walking pose have different appearance.

When the parts are detected independently, the existence of a human object can be confirmed by validating the configuration made by the detected parts. For example, in [25, 33, 38], the appearance of parts was captured by visual codewords of a dictionary while the configuration of parts was modelled in a star-like structure of parts' locations. However, the star-like structure weakly captures the regularity of parts in their poses and viewpoints. For example, it is impossible to have a valid human object whose one leg refers to a standing pose and the other leg is in a running pose. In [38], the human body was expressed in an ordered sequence of parts. The parts were considered as letters in an alphabet whereas poses were treated as words. The validation of parts was then converted to string matching in text recognition. However, modelling 2D/3D human poses in 1D sequences may omit high-ordered important information of the body layout which cannot be captured by sequences, e.g. the spatial relationship and co-occurrence of parts that are not adjacent in the 1D sequence.

In [7], Bourdev and Malik introduced a notion called "poselet". A poselet is an atomic element that represents a part in some pose and viewpoint. Unlike the conventional concept of part, a poselet may not correspond to any semantical body part. In addition, a body part in different poses and viewpoints can be represented by multiple poselets. Poselets have also found many applications such as action recognition [26], object segmentation [9] and been extended in the later works. For example, Gkioxari et al. [19] combined the DPM and poselets in the so-called $k-$poselets model in which each part in the DPM (of k parts) was represented by a poselet. In [8, 20], deep neural networks were used to learn features describing the poselets.

In general, the current use of poselets has drawbacks. In particular, the presence of a human object is verified based on individual poselets, e.g. [7]. In [6], the consistency of pairs of poselets was considered. Detected poselets were fused to form a human object in which low score detected poselets were suppressed by high score ones. However, this method implicitly assumes that the poselet detectors could locate parts perfectly, i.e. the higher the detection score is, the more accurately the poselet is detected. Meanwhile, false alarms may have higher detection scores than true detections and thus true detections may be dismissed.

1.2. Contributions

This paper aims to develop a robust part-based model that is able to handle the high articulation of human objects, enriched with the fine-grained information of the human body structure, and efficient for computation. To this end, we make the following contributions.

- We propose a so-called Markov Random Field (MRF)-poselets model using poselets for describing human parts and MRF for modelling the spatial and structural relationship between parts. We note that MRF was also used in [1] for modelling object structures. However, our model differs from the one in [1] in two points. Firstly, the observation nodes in our MRF model are not part types. Instead, they are the detected poselets (or parts) and hence, for each poselet type, more than one poselet instance can be detected in an object hypothesis. As will be explained later in section 3, this makes our model less sensitive to false poselet detections accidentally generated by poselet detectors. Secondly, our MRF model is not a tree and thus would be more robust to model various deformation modes.

- We propose a new formulation of human detection via maximum a posteriori (MAP) estimation in the MRF-poselets model.

- We propose an efficient inference method for MAP estimation using variational mean field approximation.

Compared with the DPM [15], the MRF-poselets model holds several advantages. Firstly, it does not require all part types (i.e. poselet types such as head, arms, legs, etc.) to appear in a human object and hence the number of detected parts is not fixed. This enables the model to deal with occlusions. Secondly, it is adaptive to heavy deformations of the human body given a sufficient number of poselet types. Thirdly, poselets capture both the appearance and structure of parts, thus they offer fine-grained and richer information than the deformable parts in the DPM and visual codewords in [38]. Fourthly, the MRF-poselets model is flexible and extendable to accommodate various part detectors (e.g. deep poselets [8] could be used as part detectors).

The remainder of the paper is organised as follows. Section 2 briefly presents the notion of poselet and related aspects such as poselet learning and detection. Section 3 describes the MRF-poselets model and formulates the problem of human detection. Section 4 describes the variational mean field approximation method which is applied for the MAP inference. Section 5 presents learning parameters. Experimental results and comparisons are presented in section 6. Section 7 concludes the paper with remarks.

2. Poselet

Bourdev and Malik [7] define a poselet as a part of one's pose. A poselet is associated with an appearance model and a set of keypoints (e.g. the body joints, eyes, ears, and nose) capturing the structure of the poselet. The appearance of poselets can be represented by features determined in advance (e.g. HOG [11]) or selected using deep neural networks [8, 20]. Ideally, a poselet is informative and representative for a part in some particular pose and viewpoint. However, it does not necessarily represent a semantical body part. Fig. 1 shows some poselet samples.

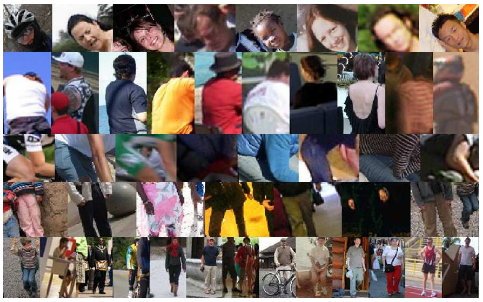


Figure 1. Some examples of poselets. As can be seen, a poselet can be just a body part (e.g. the face) or a full human body.

To describe the human body in various poses and viewpoints, poselet types are collected as follows. Given a set of training images associated with annotated keypoints, on those images, a number of image patches are sampled. For each sampled patch, other image patches having similar keypoint configuration are extracted. Those patches form a cluster and such clusters with sufficient number of members are considered as poselet types. Image patches of a cluster are used as positive samples to train a poselet detector representative for that cluster (i.e. poselet type) while patches belonging to other clusters or containing the background are considered as negative samples.

Training poselet detectors can be done in a similar manner to the work in [11]. In particular, for each poselet type, positive and negative samples are encoded with features (e.g. HOG) and used to train a classifier (e.g. linear SVM). Bootstrapping is also conducted to enhance the poselet detectors with hard false positives. Since the poselet types are selected to be tightly clustered in the configuration space of keypoints, the spatial distribution of each keypoint in relative to the location of a poselet is modelled using a Gaussian. The difference between the work in [7] and [6] is that only 2D information of keypoints is used in [6].

Detecting poselets in an input image can be done as follows. Each poselet detector scans the input image at various scales and locations. For a poselet candidate, the features are extracted and the trained classifier is invoked to classify the candidate. Candidates whose the classification score is lower than a detection threshold are filtered out. Non-maximal suppression is then used to merge nearby poselet candidates of the same poselet type (i.e. generated by the same poselet detector) to form a set of poselet detections.

3. Problem Formulation

Assume that we are given a set of poselet types $\mathcal{T} = \{t_1, ..., t_K\}$ determined as in section 2. Let $\mathcal{S} = \{s_1, ..., s_N\}$ be a set of poselets detected on an object hypothesis. Detecting poselets is also presented in section 2. Let $g : \mathcal{S} \rightarrow \mathcal{T}$ be a poselet type mapping defined as, for each $s_i \in \mathcal{S}$, $g(s_i) = t_j \in \mathcal{T}$. The mapping g is associated with an index mapping $f : \{1, ..., N\} \rightarrow \{1, ..., K\}$ defined as, for each $i \in \{1, ..., N\}$, $f(i) \in \{1, ..., K\}$ such that $g(s_i) = t_{f(i)}$, i.e. $t_{f(i)} \in \mathcal{T}$ is the poselet type fired at $s_i \in \mathcal{S}$. Note that, it is possible to have $i \neq j$ but $f(i) = f(j)$ since we allow one poselet type to be detected more than once in an object hypothesis instead of suppressing low score detected hypotheses by high score ones as in [6]. This prevents the rejection of true poselets due to the imperfectness of the poselet detectors.

The MRF-poselets model is constructed as follows. Each detected poselet $s_i \in \mathcal{S}$ is represented by a node in the model and associated with a label node l_i taking value in $\{0, 1\}$; 0 indicates a false alarm and 1 represents a true detection. A poselet s_i is connected only to its label node l_i while there are connections between label nodes. Two label nodes l_i and l_j are linked if the poselet types $t_{f(i)}$ and $t_{f(j)}$ share (predict) some common keypoints. For example, a face poselet and a head-shoulders poselet of a human object share the same the eyes and nose keypoints. However, a head-shoulders poselet and a leg poselet do not have common keypoints but they can predict the same hip keypoint in some poses. The MRF-poselets model is equivalent to a conventional two-layer MRF model (as shown in Fig. 2(a)) in which s_i are observation nodes and l_i are hidden nodes.

Similarly to the two-layer MRF model, we assume that the prior $p(\mathcal{L} = \{l_1, ..., l_N\} \in \{0, 1\}^N)$ is a Gibbs distribution, i.e. $p(\mathcal{L})$ can be factorised as,

$$p(\mathcal{L}) = \frac{1}{Z} \prod_{l_i, l_j} \psi_{ij}(l_i, l_j) \prod_{l_i} \psi_i(l_i) \tag{1}$$

where Z is a normalisation factor.

Since l_i are binary variables, we can write $p(\mathcal{L})$ in the form of a Boltzmann distribution as,

$$p(\mathcal{L}) = \frac{1}{Z} \prod_{l_i, l_j} e^{m_{f(i)f(j)} h_{f(i)f(j)} l_i l_j} \prod_{l_i} e^{n_{f(i)} l_i} \tag{2}$$

where $m_{f(i)f(j)}$ and $n_{f(i)}$ are the parameters of the MRF-poselets model and $h_{f(i)f(j)}$ is some measure of the consistency between two poselet types $t_{f(i)}$ and $t_{f(j)}$. The terms $h_{f(i)f(j)}$ are added to augment the spatial configuration information of poselets. In particular, we define,

$$h_{f(i)f(j)} = \frac{-1}{1 + e^{-(d_{f(i)f(j)} - \epsilon)}} + 0.5 \tag{3}$$

where $d_{f(i)f(j)}$ is a measure of the divergence of keypoint distributions of poselet types $t_{f(i)}$ and $t_{f(j)}$. Following [6], $d_{f(i)f(j)}$ is computed as,

$$d_{f(i)f(j)} = \frac{1}{X} \sum_x KL(\mathcal{N}^x_{t_{f(i)}} || \mathcal{N}^x_{t_{f(j)}}) + KL(\mathcal{N}^x_{t_{f(j)}} || \mathcal{N}^x_{t_{f(i)}}) \tag{4}$$

where X is the number of common keypoints shared/predicted by both poselet types $t_{f(i)}$ and $t_{f(j)}$, $\mathcal{N}^x_{t_{f(i)}}$ is the empirical distribution of a keypoint x in $t_{f(i)}$, and KL is the Kullback-Leibler divergence. Note that $\mathcal{N}^x_{t_{f(i)}}$ is computed by translating the empirical distribution of x to the location where $t_{f(i)}$ is detected. The empirical distributions $\mathcal{N}^x_{t_{f(i)}}$ are modelled by Gaussians.

It is expected that two consistent poselet types have similar keypoints distributions and thus $d_{f(i)f(j)}$ is small. In (3), the consistency between two poselet types $t_{f(i)}$ and $t_{f(j)}$ is modelled in a logistic-like function of $d_{f(i)f(j)}$. This reflects the fact that $e^{m_{f(i)f(j)}h_{f(i)f(j)}l_i l_j}$ (used in (2)) fluctuates around 1 subject to the variation of $d_{f(i)f(j)}$ around ϵ (a predefined value). The greater $d_{f(i)f(j)}$ is, the lower $e^{m_{f(i)f(j)}h_{f(i)f(j)}l_i l_j}$ is, and vice versa.

In (2), $m_{f(i)f(j)}$ represents the correlation between two different poselet types $t_{f(i)}$ and $t_{f(j)}$. It is used to augment the consistency between poselet types. Note that $d_{f(i)f(j)}$ is also used for this purpose. However, $d_{f(i)f(j)}$ is averaged over all keypoints and thus the quantity of keypoints shared by two poselet types is not considered. Meanwhile, the more keypoints that can be shared by two poselet types, the higher correlation those poselet types should have. For example, a pair of the head-shoulders and upper body of a human object shows stronger correlation than a pair of the head-shoulders and legs. This is because the former shares more common keypoints than the latter. The parameter $n_{f(i)}$ represents the importance of poselet type $t_{f(i)}$. For example, in most cases the torso poselet appears in a true human object and thus would have higher importance compared with other poselet types, e.g. legs. In our method, both $m_{f(i)f(j)}$ and $n_{f(i)}$ can be automatically estimated from the training data.

The likelihood $p(\mathcal{S}|\mathcal{L})$ is defined as,

$$p(\mathcal{S}|\mathcal{L}) = \prod_{i=1}^{N} p(s_i|l_i) \tag{5}$$

More specifically, we define,

$$p(s_i|l_i) \propto \frac{1}{1 + e^{-\alpha_{f(i)}(c_i - \beta_{f(i)})}} \tag{6}$$

where c_i is the detection score of s_i generated by a classifier (e.g. SVM), $\alpha_{f(i)}$ and $\beta_{f(i)}$ are parameters used to convert c_i to a probability [32, 39]. The parameters $\alpha_{f(i)}$ and $\beta_{f(i)}$ are learned by fitting a logistic over positive and negative poselet detections from the training dataset.

Given the model parameters $(m_{ij}, n_i), i \in \{1, ..., K\}$, the set of predefined poselet types $\mathcal{T}$ and detected poselets $\mathcal{S}$, the presence of a human object can be verified by finding the optimal $\mathcal{L}^*$ such that

$$\mathcal{L}^* = \underset{\mathcal{L}\in\{0,1\}^N}{\arg\max}\, p(\mathcal{L}|\mathcal{S}) \propto \underset{\mathcal{L}\in\{0,1\}^N}{\arg\max} \prod_{i=1}^{N} p(s_i|l_i)p(\mathcal{L}) \tag{7}$$

where $p(s_i|l_i)$ and $p(\mathcal{L})$ are defined in (6) and (2).

Since the MRF-poselets model can have cycles, the MAP inference in (7) cannot be solved by using exact inference methods (e.g. [17, 24]). In addition, a brute-force inference would be intractable due to exponential complexity. Note that, since we allow many detections of the same poselet type by not suppressing low score detections by high score detections as in [6], the number of poselet detections to be considered for a human object is quite large. To overcome this issue, variational approach is adopted in the paper.

4. Variational Inference

Variational methods are often used when exact solutions are not feasible/practical to be obtained [22, 23]. In Computer Vision, variational methods have been employed to solve various tasks such as pedestrian detection [36, 30], object tracking [28], template matching [29]. For our problem, instead of estimating $p(\mathcal{L}|\mathcal{S})$, we approximate it by a variational distribution $Q(\mathcal{L})$ which is simpler and more efficient to be computed. As shown in [35], the variational distribution $Q(\mathcal{L})$ can be found by maximising an objective function $J(Q)$ as,

$$\begin{aligned} J(Q) &= \log p(\mathcal{S}) - KL(Q(\mathcal{L})||p(\mathcal{L}|\mathcal{S})) \\ &= -\int_{\mathcal{L}} Q(\mathcal{L}) \log Q(\mathcal{L}) d\mathcal{L} + \int_{\mathcal{L}} Q(\mathcal{L}) \log p(\mathcal{L}, \mathcal{S}) d\mathcal{L} \\ &= \mathcal{H}(Q) + E_Q\{\log p(\mathcal{L}, \mathcal{S})\} \end{aligned} \tag{8}$$

where $\mathcal{H}(Q)$ is the entropy of the variational distribution Q, $E_Q\{\cdot\}$ represents the expectation with regard to Q. For the sake of simplicity, Q shall be used to replace $Q(\mathcal{L})$ when there is no ambiguity.

As shown in (8), since KL is non-negative, maximising $J(Q)$ will result in an approximation of $p(\mathcal{L}|\mathcal{S})$. The most important factor in variational methods is to select the variational distribution Q. In this paper, the simplest variational distribution of Q assuming a full factorisation is used, i.e.

$$Q(\mathcal{L}) = \prod_{i=1}^{|\mathcal{L}|} Q_i(l_i) \tag{9}$$

where $Q_i(l_i)$ is the variational distribution of l_i.

Since l_i are binary variables, we specifically define,

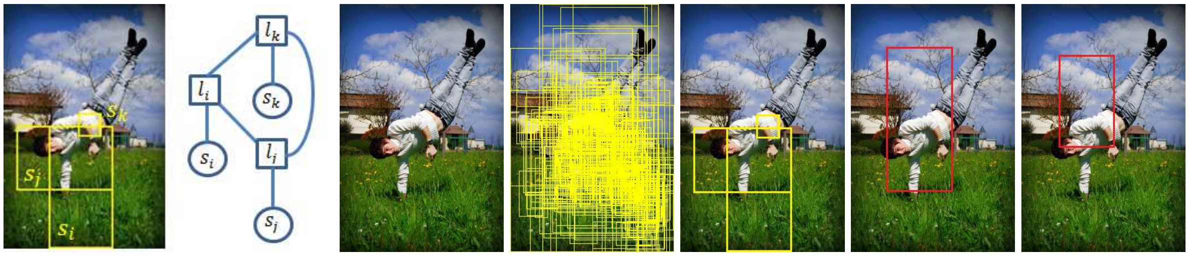
$$Q_i(l_i) = \mu_i^{l_i}(1 - \mu_i)^{(1-l_i)} \tag{10}$$

where μ_i is computed as,

$$\mu_i = \frac{p(s_i|l_i = 1)k_i}{p(s_i|l_i = 0) + p(s_i|l_i = 1)k_i} \tag{11}$$

where $p(s_i|l_i)$ is defined in (6) and

$$k_i = \exp\left\{ \sum_{l_j \in \mathcal{V}(l_i)} m_{f(i)f(j)} h_{f(i)f(j)} \mu_j + n_{f(i)} \right\} \tag{12}$$



(a) (b)

Figure 2. (a) An MRF-poselets with three poselet detections. (b) From left to right: Input image, poselet detections, poselets whose label is 1, our detection result obtained by applying the bounds prediction method in [6], result of [6] obtained by fusing all detected poselets.

where $\mathcal{V}(l_i)$ denotes the set of label nodes whose the corresponding poselet type and $t_{f(i)}$ share/predict some common keypoints.

As shown in (11) and (12), μ_i is updated locally based on the neighbouring nodes in $\mathcal{V}(l_i)$ and the update is performed iteratively to increase $J(Q)$ which is finally computed as,

$$\begin{aligned} J(Q) = &\sum_i \mathcal{H}(Q_i) + \sum_{i,j} m_{f(i)f(j)} h_{f(i)f(j)} \mu_i \mu_j \\ &+ \sum_i n_{f(i)} \mu_i + \sum_i (1-\mu_i) \log p(s_i | l_i = 0) \\ &+ \sum_i \mu_i \log p(s_i | l_i = 1) - \log Z \end{aligned} \tag{13}$$

where $\mathcal{H}(Q_i)$ is the entropy of the individual variational distribution Q_i and $\mathcal{H}(Q) = \sum_i \mathcal{H}(Q_i)$ due to the full factorisation of Q.

The estimation of $J(Q)$, as shown in (13), requires the computation of Z, which again takes an exponential complexity. However, the optimisation of $J(Q)$ can be done without involving Z by using an alternative objective function $\widetilde{J}(Q) = J(Q) + \log Z$. Once the optimal variational distribution Q^* has been obtained, it can be used to approximate $p(\mathcal{L}|\mathcal{S})$. In particular, since Q is fully factorised, we can approximate

$$p(\mathcal{L}^* | \mathcal{S}) \approx \prod_i^{|\mathcal{L}|} Q_i(l_i^*) \tag{14}$$

where $l_i^* = \arg\max_{l_i} Q_i(l_i)$.

Given the optimal labelling configuration $\mathcal{L}^*$ obtained by the variational inference, the method in [6] can be used to predict the human object bounds (i.e. the bounding boxes). However, the prediction is applied only to poselet activations s_i whose the optimal label $l_i^* = 1$. In particular, each detected poselet votes for the bounds of a candidate human object containing that poselet. The bounds are then clustered using mean shift algorithm. Fig. 2(b) illustrates an example of the detection method.

5. Learning Parameters

In this section, we present how to learn parameters $(m_{ij}, n_i), i \in \{1, ..., K\}$. The learning process is conducted in a similar way to the expectation-maximisation (EM) fashion [2]. In particular, let $\mathcal{B}$ be a set of annotated human objects and $\mathcal{S}_b = \{s_1, ..., s_{N_b}\}$ be the set of poselets detected on a human object $b \in \mathcal{B}$. For each poselet type index $i \in \{1, ..., K\}$, we define an inverse mapping of f on b as $f_b^{-1}(i) = \{j \in \{1, ..., N_b\} | g(s_j) = t_i\}$, i.e. $f_b^{-1}(i)$ is the set of indices of detected poselets that correspond to poselet type t_i. The learning process is conducted iteratively in two steps as follows.

- **E-step:** Given a setting of $(m_{ij}, n_i), i \in \{1, ..., K\}$ and the set of annotated human objects $\mathcal{B}$, poselets are detected and MAP estimations $p(\mathcal{L}^* | \mathcal{S}_b)$ are approximated as in (14) using the variational inference algorithm presented in section 4. In other words, variational parameters $\mu_i, i \in \{1, ..., N_b\}$ and distributions $Q_i, i \in \{1, ..., N_b\}$ are determined for each $b \in \mathcal{B}$.
- **M-step:** Given $\mu_i, Q_i, i \in \{1, ..., N_b\}$ for each $b \in \mathcal{B}$, we define a general objective function $J_{\mathcal{B}}(Q) = \sum_{b \in \mathcal{B}} J_b(Q)$ where each $J_b(Q)$ is optimised on a human object b using (13). The parameters $(m_{ij}, n_i), i \in \{1, ..., K\}$ are then updated using a similar manner to the gradient descent method in which the increments are set proportionally to $\frac{\partial J_{\mathcal{B}}(Q)}{\partial m_{ij}}$ and $\frac{\partial J_{\mathcal{B}}(Q)}{\partial n_i}$ as,

$$\frac{\partial J_{\mathcal{B}}(Q)}{\partial m_{ij}} = m_{ij} \sum_{b \in \mathcal{B}} \sum_{u \in f_b^{-1}(i), v \in f_b^{-1}(j)} h_{ij} \mu_u \mu_v \tag{15}$$

$$\frac{\partial J_{\mathcal{B}}(Q)}{\partial n_i} = n_i \sum_{b \in \mathcal{B}} \sum_{u \in f_b^{-1}(i)} \mu_u \tag{16}$$

6. Experimental Results

6.1. Experimental Setup

We re-used the pre-trained poselet detectors [6] publicly available (for both poselet selection and training poselet detectors). There were 150 poselet detectors used in our experiments. Our training phase mainly aimed to learn parameters $(m_{ij}, n_i), i \in \{1, ..., K\}$ (as presented in Section 5). The training of the poselet detectors was conducted on the PASCAL VOC 2009 training set.

For the construction of the MRF-poselets models, poselets are detected once on the whole input image using the pre-trained poselet detectors in [6]. Two label nodes l_i and l_j are connected if $d_{f(i)f(j)} < \tau$ where $d_{f(i)f(j)}$ is defined in (4) and τ is a user-defined value. As provided in the poselet-based object detector [6], τ was set to 5. Based on this setting, ϵ in (3) was set to 3. All the observation and label nodes are clustered using a labelling connected component algorithm. Each connected component is considered as an MRF-poselets model on which the inference is performed.

The variational inference algorithm is iterative and stops when an optimal solution has been reached or the number of iterations has exceeded a predefined value. In our experiments, the maximum of number of iterations was set to 100. The variational distributions Q_i could be initialised as $Q_i(l_i = 1) = 1 - Q_i(l_i = 0) = 0.5$. We found that, the detection accuracy was improved when $Q_i(l_i = 1)$ were set proportionally to the detection score of s_i.

6.2. Evaluation, Comparison, and Analysis

We evaluated the proposed MRF-poselets model and compared it with existing methods on three tasks: torso detection, person detection, and keypoint prediction. The proposed model was evaluated on different datasets including H3D [7] and PASCAL VOC 2007-2009 [14]. The detection performance was measured using the precision-recall (PR) and average precision (AP) metric. True detections and false alarms were determined by using the PASCAL VOC criterion [14]. Specifically, a detected object is considered as true detection if there exists a ground-truth object that matches the detected object. A match is confirmed if the ratio of the intersection and union of the bounding boxes covering these two objects is greater than 0.5.

Torso Detection. We note that torso detection is important and worthy to be used for evaluating part-based models. This is because the human torso is more stable compared with the entire body's bounding box. In addition, amongst the body parts, the torso is more often found even under partial occlusions. The torso can be predicted based on the locations of the hips and shoulders which are predicted by averaging the prediction of detected poselets. The detection score of each detected poselet is also used in the prediction. The aspect ratio of the torso is fixed to 1.5.

This experiment was conducted on the test set of the H3D dataset and the validation set of the PASCAL VOC 2009. On the H3D dataset, the information of keypoints is available for the evaluation. On the validation set of the PASCAL VOC 2009, the annotations of keypoints are provided publicly in [19]. The detection performance of the proposed method and existing methods on the H3D test set and PASCAL VOC 2009 validation set are presented in Fig. 3 and Table 1 respectively. Some examples of our method are shown in Fig. 4.

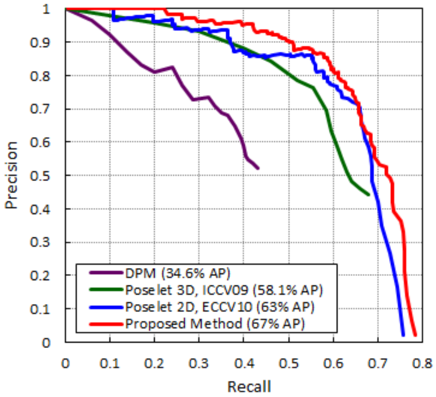


Figure 3. Performance of torso detection on H3D dataset.

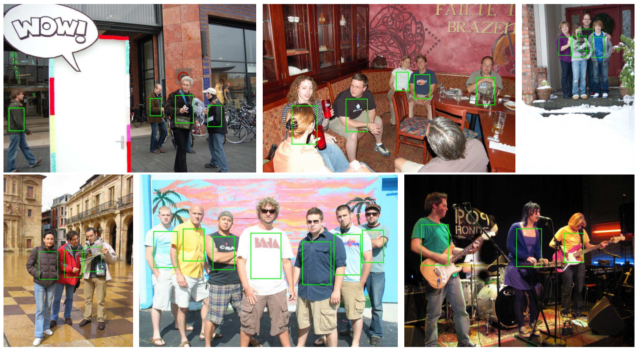


Figure 4. Example detection results on the H3D dataset.

Table 1. Average Precision (AP) of torso detection evaluated on the second half of the PASCAL VOC 2009 validation set.

Method	AP
DPM [15]	21.9
Poselet 2D [6]	26.1
1-poselets [19]	27.6
$\{1, 2\}$-poselets [19]	27.8
MRF-poselets	31.7

As shown in Table 1, the proposed method significantly outperforms the original works of poselets [7, 6] and achieves superior performance in comparison to the state-of-the-art on torso detection. Our method improves both the precision and recall rate. Specifically, on the H3D dataset, our method increases the AP approximately by 4% compared with [6], 9% compared with [7], and 12% higher than that of the DPM [16]. On the PASCAL VOC 2009 dataset, our method obtains the best performance, about 4% higher AP compared with the current state-of-the-art [19].

Person Detection. For evaluation of person detection, we tested and compared the proposed method with existing methods on the test sets of the PASCAL VOC 2007-2009. The detection results of our method and other methods are reported in Table 2. Several detection results are shown in Fig. 5. As can be seen from Fig. 5, the proposed method can robustly detect humans in various poses and viewpoints (the 1st and 2nd row of Fig. 5) and potentially deal with partial occlusions (the 3rd and 4th row of Fig. 5).

Compared with torso detection, the performance of our person detection was comparable to that of the poselets method in [6]. In particular, our method obtained a better AP on the PASCAL VOC 2007 but incurred a lower precision on the PASCAL VOC 2009. Through experiments, we have found that this is because the prediction of the human object bounds is sensitive to occlusions in which many of the poselets detected on occluded parts are filtered out by the inference algorithm. We also noticed that methods using deep learning, e.g. the deep poselets [8], obtained superior performance on the PASCAL VOC 2007 dataset. We consider using deep poselets to substitute the current HOG-poselets to enhance the performance of our model as the future work. Note that such substitution is possible because the proposed MRF-poselets model is adaptive to various part detectors.

Table 2. Average Precision (%) of the proposed method and existing methods for person detection task.

Method	VOC 2007	VOC 2009
DPM [16]	36.8	
DPM v5 [15]	43.2	43.8
Poselet 3D [7]	36.5	
Poselet 2D [6]	46.5	47.8
Boosted HOG-LBP [40]	44.6	
1-poselets [19]	45.6	
$\{1, 2\}$-poselets [19]	45.4	
R-CNN [18]	58.7	
Deep poselets [8]	59.3	
MRF-poselets	47.6	47.1

Keypoint Prediction. Compared with other notions of parts, e.g. the deformable parts in the DPM [15], visual codewords in [38], poselets are richer with the structure information represented by the keypoints. In our experiment, keypoint prediction was conducted as follows. For each human hypothesis, an MRF-poselets model was constructed and the inference algorithm was applied. After the inference, each keypoint in the human object could be predicted by more than one poselet detection and only the poselet detection s_i with the maximum of $Q_i(l_i^* = 1)$ was used to compute the keypoint. Fig. 6 shows several results of keypoint prediction.

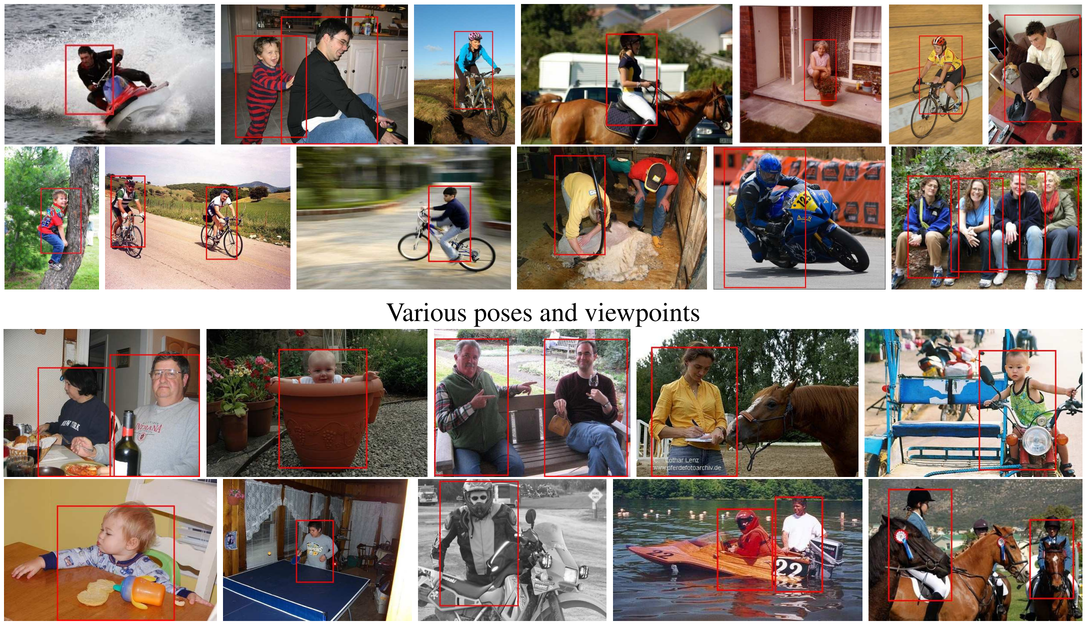
The performance of keypoint prediction was measured using the average precision of keypoints (APK) metric proposed in [37]. The APK is similar to the AP used for evaluation of torso/person detection. Specifically, a predicted keypoint is considered as true prediction if its Euclidean distance to the ground truth keypoint is not over $\alpha \cdot h$ where h is the height of the torso in the ground truth. Similarly to [19], α was set 0.2 in our evaluation. Table 3 compares the MRF-poselets model with other models for keypoint prediction on the validation set of the PASCAL VOC 2009. As shown in Table 3, the MRF-poselets model achieves the best overall performance. It also performs best in most cases. We also notice that our model significantly outperforms the current state-of-the-art [20] using convolutional neural networks for keypoint prediction on the PASCAL VOC 2009 validation set. This again shows that our model could potentially gain further improvement when deep learning, e.g. deep poselets [8], is used.

Computational Analysis. We also investigated the complexity of the MRF-poselets model based on the poselet detection time, the processing time and number of iterations required for the inference, and the overall time to process an image. This experiment was conducted on the PASCAL VOC 2007 dataset and on an Intel(R) Core(TM) i7 2.10GHz CPU computer with 8.00 GB memory. Table 4 presents the computational measures. As shown in Table 4, our method could process an image in about 15.41 seconds in which the time spent for detecting poselets was approximately 15 seconds while the inference algorithm took just 0.011 seconds. We have also found that with the poselet-based object detector [6], the processing time spent for fusing all the detected poselets to form object hypotheses was about 3.5% of the overall processing time, i.e. higher than our inference algorithm. This is probably because the inference algorithm could significantly filter out false poselet detections (i.e. poselet detections whose label is 0). On the average, the inference on each hypothesis was done in about 4 iterations while the maximum number of interations was about 10 times to reach the optimum.

Mean Field Inference Analysis. Variational mean field inference is a local optimisation method and its quality can be evaluated via the difference $\log p(\mathcal{S}) - J(Q)$ or the tightness of $\{Q_i(l_i)\}$ to $p(l_i|\mathcal{S})$. However, since the true labels of l_i are not available, this quality is not evaluated directly but indirectly via the detection/prediction accuracy. As show in [22], graphs with weak dependencies between nodes are expected to have good approximate. In our case, we have found that the presence of a poselet is affected only by its nearby poselets. This explains the success of the mean field inference in the experiments.

Table 4. Computational complexity of the MRF-poselets model for human detection. All measures are referred to one image and computed by averaging all images in the PASCAL VOC 2007 dataset.

Measure	Complexity
Poselet detection time	15.10 (seconds)
Inference time	0.01 (seconds)
Overall	15.41 (seconds)
#Iterations	4.02
#Groups	131



Various poses and viewpoints

Under occlusions

Figure 5. Example results of person detection.

Table 3. Performance evaluation using average precision of keypoints (APK).

	Nose	R_Shoulder	R_Elbow	R_Wrist	L_Shoulder	L_Elbow	L_Wrist	R_Hip	R_Knee	R_Ankle	L_Hip	L_Knee	L_Ankle	AVG
Poselets [6]	22.5	16.0	6.4	4.3	16.1	5.8	3.2	7.5	2.0	2.8	7.6	3.1	6.6	8.0
DPM [15]	39.1	25.2	7.7	1.8	25.1	7.8	1.3	8.5	1.3	1.0	9.7	1.5	1.3	10.1
Yang's model [37]	-	13.7	7.7	4.3	15.5	8.9	5.2	3.0	3.3	3.9	4.0	3.1	4.1	-
1-poselets [19]	44.9	24.9	9.7	3.4	26.4	10.9	3.0	9.0	2.5	3.0	8.4	2.8	2.2	11.6
$\{1, 2\}$-poselets [19]	42.9	27.1	12.2	3.4	27.3	11.8	2.8	10.6	4.4	3.8	11.4	4.9	3.2	12.7
R-CNN [20]	**52.0**	32.5	**16.6**	**5.9**	32.1	14.6	5.6	9.7	4.0	4.6	10.8	4.8	4.8	15.2
MRF-poselets	50.3	**41.7**	15.0	5.1	**39.8**	**15.9**	**6.6**	**24.5**	**4.8**	**4.8**	**23.1**	**6.7**	**8.0**	**19.0**

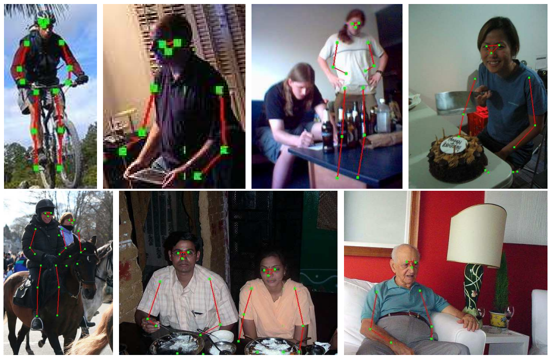


Figure 6. Keypoint prediction on the VOC 2009 dataset.

7. Conclusion

This paper proposes a so-called MRF-poselets model constructed from poselets for representing parts and MRF for modelling the human body structure. The task of human detection is formulated as a MAP estimation and efficiently solved using variational mean field inference. The proposed model was verified and compared with existing methods in various tasks including torso detection, person detection, and keypoint prediction. Experimental results have shown that the model can be applied in detecting humans in various poses, viewpoints and under occlusions and achieves the state-of-the-art performance on both torso detection and keypoint prediction. The proposed model can accommodate various part detectors and is being improved by cooperating with deep neural networks as part detectors.

8. Acknowledgement

Sai-Kit Yeung is supported by SUTD-ZJU Collaboration Research Grant 2012 SUTDZJU/RES/03/2012, SUTD-MIT International Design Center Grant IDG31300106, and Singapore MOE Academic Research Fund MOE2013-T2-1-159. Duc Thanh Nguyen and Minh-Khoi Tran are supported by Singapore MOE Academic Research Fund MOE2013-T2-1-159. We also acknowledge the support of the SUTD Digital Manufacturing and Design (DManD) Centre which is supported by the Singapore National Research Foundation.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009. 1, 2

[2] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003. 5

[3] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool. Pedestrian detection at 100 frames per second. In *CVPR*, pages 2903–2910, 2012. 1

[4] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool. Seeking the strongest rigid detector. In *CVPR*, pages 3666–3673, 2013. 1

[5] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*, 2014. 1

[6] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, pages 168–181, 2010. 1, 2, 3, 4, 5, 6, 7, 8

[7] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, pages 1365–1372, 2009. 1, 2, 3, 6, 7

[8] L. D. Bourdev, F. Yang, and R. Fergus. Deep poselets for human detection. *CoRR*, abs/1407.0717, 2014. 1, 2, 7

[9] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, pages 2225–2232, 2011. 2

[10] A. D. Costea and S. Nedevschi. Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In *CVPR*, pages 2393–2400, 2014. 1

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 1, 2, 3

[12] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, pages 1–14, 2014. 1

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012. 1

[14] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 6

[15] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 1, 2, 6, 7, 8

[16] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8, 2008. 1, 6, 7

[17] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 4

[18] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 1, 7

[19] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using *k*-poselets for detecting people and localizing their keypoints. In *CVPR*, pages 3582–3589, 2014. 1, 2, 6, 7, 8

[20] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-cnns for pose estimation and action detection. 2015. 1, 2, 7, 8

[21] S. Hussain and B. Triggs. Feature sets and dimensionality reduction for visual object detection. In *BMVC*, pages 1–10, 2010. 1

[22] T. S. Jaakkola. Tutorial on variational approximation methods. Technical report, MIT Artificial Intelligence Laboratory, 2000. 4, 7

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, pages 183–233, 1999. 4

[24] F. R. Kschischang, B. J. Frey, and H. A. Loelinger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47(2):498–519, 2001. 4

[25] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages 878–885, 2005. 1, 2

[26] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, pages 3177–3184, 2011. 2

[27] S. Maji and J. Malik. Object detection using a max-margin hough tranform. In *CVPR*, 2009. 1

[28] C. Medrano, J. E. Herrero, J. Martínez, and C. Orrite. Mean field approach for tracking similar objects. *CVIU*, 113:907–920, 2009. 4

[29] D. T. Nguyen. A novel chamfer template matching method using variational mean field. In *CVPR*, pages 2425–2432, 2014. 4

[30] D. T. Nguyen, W. Li, and P. Ogunbona. Inter-occlusion reasoning for human detection based on variational mean field. *Neurocomputing*, 110:51–61, 2013. 4

[31] D. T. Nguyen, W. Li, and P. Ogunbona. Human detection from images and videos: a survey. *Pattern Recognition*, 2015. In press. 1

[32] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 4

[33] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, pages 1582–1588, 2006. 1, 2

[34] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005. 1

[35] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. 4

[36] Y. Wu and T. Yu. A field model for human detection and tracking. *PAMI*, 28(5):753–765, 2006. 4

[37] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *PAMI*, 35(12):2878–2890, 2013. 1, 7, 8

[38] C. Yao, X. Bai, W. Liu, and L. J. Latecki. Human detection using learned part alphabet and pose dictionary. In *ECCV*, 2014. 1, 2, 7

[39] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 694–699, 2002. 4

[40] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *CVPR*, pages 1393–1400, 2011. 1, 7