PoseField: An Efficient Mean-field based Method for Joint Estimation of Human Pose, Segmentation, and Depth

Vibhav Vineet*, Glenn Sheasby, Jonathan Warrell and Philip H.S. Torr

{vibhav.vineet-2010, glenn.sheasby, jwarrell, philiptorr}@brookes.ac.uk
Oxford Brookes University
Oxford, UK

Abstract. Many models have been proposed to estimate human pose and segmentation by leveraging information from several sources. A standard approach is to formulate it in a dual decomposition framework. However, these models generally suffer from the problem of high computational complexity. In this work, we propose **PoseField**, a new highly efficient filter-based mean-field inference approach for jointly estimating human segmentation, pose, per-pixel body parts, and depth given stereo pairs of images. We extensively evaluate the efficiency and accuracy offered by our approach on H2View [1], and Buffy [2] datasets. We achieve 20 to 70 times speedup compared to the current state-of-the-art methods, as well as achieving better accuracy in all these cases.

1 Introduction

Human pose estimation and segmentation have long been popular tasks in computer vision, and a large body of research has been developed on these problems [3–7]. Several of these methods model pose estimation and segmentation problems separately, and fail to capture the large variability and deformation in appearance and the structure of humans.

However, when segmentation and pose estimation results are considered together, one can observe discrepancies, for example a foreground region not corresponding to any detected body part, or vice versa. Joining the two problems together, either sequentially or simultaneously, can help to remove these discrepancies. Researchers have thus begun to consider the possibility of jointly estimating these outputs, leveraging the information from several high-level and low-level cues.

A number of methods insert various algorithms into a pipeline, where the result of one algorithm is used to initialize another. For example, Bray et al. tackle the problem of human segmentation by introducing a pose-specific MRF, encouraging the segmentation result to look "human-like" [8]. Similarly, Kumar et al. use layered pictorial structures to generate an object category specific MRF

* The first two authors contributed to this work equally as joint first author.

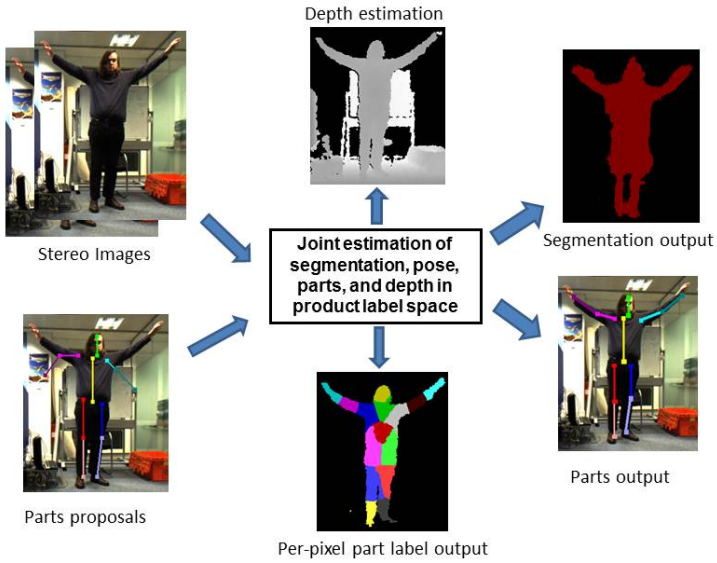

Fig. 1. Given stereo pairs and initial part proposals, our approach jointly estimates the human segmentation, pose, and depth, considering the relationships between per-pixel body part labels and part configurations.

to improve segmentation [9]. The problem with this kind of approach is that errors in one part of the algorithm can propagate to later stages. Joint inference can be used to overcome this issue; Ladický et al. obtain an improvement in object class segmentation by incorporating global information from object detectors and object co-occurrence terms [10], solving detection and segmentation with one CRF. Further, Ladický et al. frame joint estimation of object classes and disparity as CRF problems in the product label space, and solve the two tasks together [11].

Additionally, in the context of human pose estimation and segmentation, Wang and Koller propose a dual-decomposition based inference method [12] in a multi-level CRF framework to jointly estimate pose and segmentation by introducing variables that capture the coupling between these two problems [13]. Extending their formulation, Sheasby et al. [1] add depth information, thus allowing human pose, segmentation and depth to be solved together [14, 1].

The complexity of such joint frameworks is a serious issue; if the framework is to be used for applications such as security and video gaming, fast output is required. In such situations, it might prove desirable to find an efficiently solvable approximation of the original problem. One such method that can be applied here is mean-field inference [15]. For a certain class of pairwise terms, mean-field inference has been shown to be very powerful in solving the object class segmentation problem, and object-stereo correspondence problems in CRF frameworks, providing an order-of-magnitude speedup [16]. In this paper, we

propose a highly efficient filter-based mean-field approach to perform joint estimation of human segmentation, pose, per-pixel part labels, and disparity in the product label space, producing a significant improvement in speed.

Further, to model the human skeleton, we propose a hierarchical model that captures relations on multiple levels. At the lowest level, we estimate part labels per pixel. Such a representation has been shown to be successful in generating body parts proposals and pose estimation by Shotton et al. [17]. Secondly, the higher level tries to find the best configuration from a set of part proposals. Our framework is represented graphically in Fig 1.

Finally we extensively evaluate the efficiency and accuracy offered by our mean-field approach on two datasets: H2View [14], and Buffy [2]. We show results for segmentation, per pixel part labelling and pose estimation; disparity computation is used to improve these results, but is not quantitatively evaluated as it is not feasible to obtain dense ground truth data. We achieve 20-70 times speedup compared to the current state-of-the-art graph-cuts based dual-decomposition approach [1], as well achieving better accuracy in all cases.

The remainder of the paper is structured as follows: an overview of dense CRF formulation is introduced in the next section, while our body part formulation is discussed in Section 3. We describe our joint inference framework in Section 4 and learning of different parameters is discussed in the Section 5. Results follow in Section 6, and Section 7 concludes the paper.

2 Overview of Dense Random Field Formulation

The goal of our joint optimization framework is to estimate human segmentation and pose, together with part labels at the pixel level, and perform stereo reconstruction given a pair of stereo images. These problems however can be separately solved in a conditional random field (CRF) framework. Thus, before going into the details of the joint modelling and inference, we provide the models for solving them separately. Let $\mathcal{X}^{\mathcal{S}} = \{X_1^S, ..., X_N^S\}$, $\mathcal{X}^{\mathcal{J}} = \{X_1^J, ..., X_N^J\}$, $\mathcal{X}^{\mathcal{D}} = \{X_1^D, ..., X_N^D\}$ be the human segmentation, per-pixel part and disparity variables respectively. We assume each of these random variables is associated with each pixel in the image $\mathcal{N} = \{1, ..., N\}$. Further, each X_i^S takes a label from segmentation label set $\mathcal{L}^S \in \{0, 1\}$, X_i^D takes a label from $\mathcal{L}^D \in \{0...D\}$ disparity labels and X_i^J takes a label from $\mathcal{L}^J \in \{0, 1, ..., M\}$ where 0 represents background and M is the number of body parts.

First, we give details of the energy function for the segmentation variables. Assuming the true distribution of the segmentation variables is captured by the unary and pairwise terms, the energy function takes the following form:

$$E^S(\mathbf{x}^S) = \sum_{i \in V} \psi_u^S(x_i^S) + \sum_{i \in V, j \in N_i} \psi_p^S(x_i^S, x_j^S) \tag{1}$$

where N_i represents the neighborhood of the variable i, $\psi_u^S(x_i^S)$ represent unary terms for human segmentation class and $\psi_p^S(x_i^S, x_j^S)$ are pairwise terms capturing the interaction between a pair of segment variables. The human object specific unary cost $\psi_u^S(x_i^S)$ is computed based on a boosted unary classifier on image-specific appearance using the model of Shotton et al. [19]. The pairwise

terms between human segmentation variables ψ_p^S take the form of Potts models weighted by edge-preserving Gaussian kernels [18] as:

$$\psi_p^S(x_i^S, x_j^S) = \mu(x_i^S, x_j^S) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i, \mathbf{f}_j) \tag{2}$$

where $\mu(.,.)$ is an arbitrary *label compatibility function*, while the functions $k^{(v)}(.,.)$, $v = 1...V$ are Gaussian kernels defined on feature vectors $\mathbf{f}_i$, $\mathbf{f}_j$ derived from the image data at locations i and j (where Krahenbuhl and Koltun [18] form $\mathbf{f}_i$ by concatenating the intensity values at pixel i with the horizontal and vertical positions of pixel i in the image), and $w^{(v)}$, $m = 1...V$ are used to weight the kernels.

Similarly we define the energy functions over the per-pixel part and disparity variables as:

$$E^J(\mathbf{x^J}) = \sum_{i \in V} \psi_u^J(x_i^J) + \sum_{i \in V, j \in N_i} \psi_p^J(x_i^J, x_j^J) \tag{3}$$

$$E^D(\mathbf{x^D}) = \sum_{i \in V} \psi_u^D(x_i^D) + \sum_{i \in V, j \in N_i} \psi_p^D(x_i^D, x_j^D) \tag{4}$$

where $\psi_u^J(x_i^J)$ and $\psi_u^D(x_i^D)$ represent unary term for the per-pixel part and disparity variables respectively, and $\psi_p^J(x_i^J, x_j^J)$ and $\psi_p^D(x_i^D, x_j^D)$ are pairwise terms capturing the interaction between pairs of per-pixel part and disparity variables respectively. The per-pixel part variable dependent unary cost $\psi_u^J(x_i^J)$ is computed based on a boosted unary classifier on depth image. Further, if we do not have ground truth for the depth map, we can learn the unary cost for the per-pixel parts on image-specific appearance. The unary cost $\psi_u^D(x_i^D)$ for the disparity variables measures the color agreement of a pixel with its corresponding pixel i from the stereo-pair given a choice of disparity x_i^D. The pairwise terms for both these variables ψ_p^J and ψ_p^D take the form of contrast-sensitive Potts models as mentioned earlier.

3 Joint Formulation

The goal of our joint optimization framework is to estimate human segmentation and pose, together with part labels at the pixel level, and also perform stereo reconstruction. We formulate the problem in a conditional random field (CRF) framework as a product label space in a hierarchical framework. At the lower level, we define the random variables $\mathcal{X} = [\mathcal{X}^S, \mathcal{X}^J, \mathcal{X}^D]$, where $\mathcal{X}$ takes a label from the product label space $\mathcal{L} = \{(\mathcal{L}^S \times \mathcal{L}^J \times \mathcal{L}^D)^N\}$. For specifying human pose, we define a second layer, represented by a set of latent variables $\mathcal{Y} = \{Y_1, Y_2, ..., Y_M\}$ corresponding to the M body parts, each taking labels from $\mathcal{L}^P \in \{0, ..., K\}$ where $1, 2, ..., K$ corresponds to the K part proposals generated for each body part, and zero represents the background class. We generate K part proposals using the model of Yang and Ramanan [7]. The graphical model explaining our hierarchical joint model is shown in the Fig 2.

3.1 Joint energy function

Given the above model, we wish to define an energy function which is general enough to capture sufficient mutual interaction between the variables while still

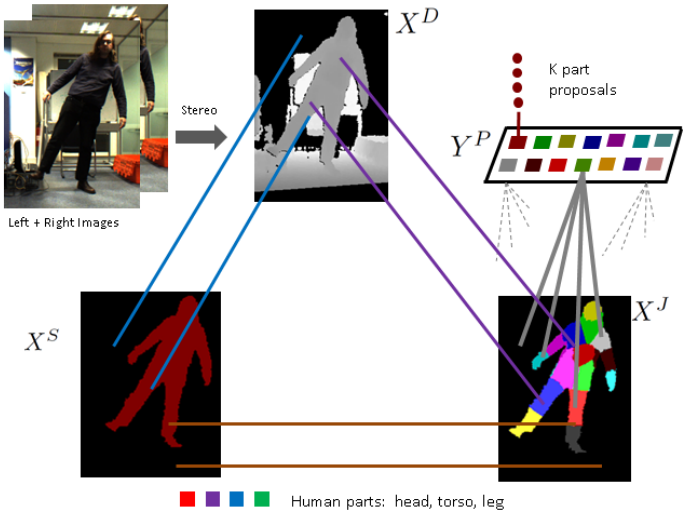

Fig. 2. PoseField model jointly estimates the per-pixel human/background segmentation, body part, and disparity labels. Further, the relationship between per pixel body part label, and part configurations are captured in a hierarchical model with information propagating between these different layers. (Best viewed in color)

providing scope for efficient inference. For this reason, we assume our energy function to take the following form:

$$E(\mathbf{x},\mathbf{y}) = E^S(\mathbf{x}^S) + E^D(\mathbf{x}^D) + E^J(\mathbf{x}^J) + E^P(\mathbf{y}) \\ + E^{SJ}(\mathbf{x}^S,\mathbf{x}^J) + E^{SD}(\mathbf{x}^S,\mathbf{x}^D) + E^{DJ}(\mathbf{x}^D,\mathbf{x}^J) + E^{JP}(\mathbf{x}^J,\mathbf{y}) \quad (5)$$

Here, our joint model has been factorized into separate layers representing human segmentation, disparity, per pixel part and latent part variables. The individual terms at the layers are captured by E^S, E^D, E^J as defined earlier and E^P, the energy function for the latent part variables, details of which are provided later in this section. Further, in order to incorporate the dependency between these variables, we add pairwise interactions between these CRF layers. E^{SJ}, E^{SD}, E^{DJ} captures the interaction between (*segment*,*per-pixel part*), (*segment*,*disparity*) and (*disparity*, *per-pixel part*) variables. The term E^{JP} captures the mutual interaction between the (*per-pixel part*, *latent part*) variables. We design the forms of these pairwise interactions to allow efficient and accurate inference; details are provided below.

Per-part terms E^P : In our hierarchical model, the top layer corresponds to the human part variables Y, which involve per-part unary cost $\psi_u^P(x_i = k)$ for associating the i^{th} part to the k^{th} proposal or to the background [1], and the pairwise term $\psi_p^P(y_i, y_j)$ penalizes the case where parts that should be connected are distant from one another in image space. The per-part unary term $\psi_u^P(y_i = k)$ is the score generated by the Yang and Ramanan model [7].

Segment, per-pixel part terms (E^{SJ}) : The joint human segmentation and per-pixel body part term, E^{SJ}, encodes the relation between segmentation and per-pixel body part. Specifically, we expect a variable that takes a body part label to belong to the foreground class, and vice versa. We pay a cost of C^{SJ} for violation of this constraint, incorporated through a pairwise interaction between the segmentation and per-pixel part variables; this interaction takes the following form:

$$E^{SJ} = \psi_p^{SJ}(\mathbf{x}^S,\mathbf{x}^J) = \sum_{i=1}^{N} C^{SJ} \cdot [(x_i^S = 1) \wedge (x_i^J = 0)] + \sum_{i=1}^{N} C^{SJ} \cdot [(x_i^S = 0) \wedge (x_i^J \neq 0)] \quad (6)$$

Segment, disparity terms (E^{SD}) : Additionally, our joint object-depth cost E^{SD} encourages pixels with a high disparity to be classed as foreground, and pixels with a low disparity to be classified as background. We penalize the violation of this constraint by a cost C^{SD}. Following the formulation of [1], we first generate a segmentation map $\mathcal{F} = \{F_1, F_2, ..., F_N\}$ by thresholding the disparity map, thus each F_i takes a label from L^S. We would expect the prior map $\mathcal{F}$ to agree with the segmentation result, so that pixels taking human label ($f_i = 1$) are classified as human, and vice versa, otherwise we pay a cost C^{SD} for violation of this constraint:

$$E^{SD} = \psi_p^{SD}(\mathbf{x}^S, \mathbf{x}^D) = \sum_{i=1}^{N} C^{SD} \cdot [(x_i^S = 1) \wedge (f_i = 0)] + \sum_{i=1}^{N} C^{SD} \cdot [(x_i^S = 0) \wedge (f_i = 1)] \quad (7)$$

Per-pixel part, disparity terms (E^{JD}) : The joint energy term E^{JD} encodes the relationship between the per-pixel body part variables and the disparity variables. As with the cost term E^{SD}, we use a flood fill to generate a segmentation map $\mathcal{F} = \{F_1, F_2, \ldots, F_N\}$ which gives us a prior based on disparity. We expect pixels classed as human by this prior (so $f_i = 1$) to be assigned to a body part label, so $x_i^J > 0$. Conversely, pixels classed as background ($f_i = 0$) should be assigned to the background label ($x_i^J = 0$). Therefore, the energy term has the following form:

$$E^{JD} = \psi_p^{JD}(\mathbf{x}^J, \mathbf{x}^D) = \sum_{i=1}^{N} C^{JD} \cdot [(x_i^J > 0) \wedge (f_i = 0)] + \sum_{i=1}^{N} C^{JD} \cdot [(x_i^J = 0) \wedge (f_i = 1)] \quad (8)$$

Per-pixel part, latent part terms (E^{JP}) : E^{JP} enforces the constraint that when a body part l is present in the solution at the pixel level, then the variable Y_l^P corresponding to the part l must be on, otherwise we pay a cost of C^{JP}.

$$E^{JP} = \psi_p^{JP}(\mathbf{x}^J, \mathbf{y}) = C^{JP} \cdot \sum_{l \in M} [(y_l = 0) \wedge (\sum_i [x_i^J = l]) > 0] \tag{9}$$

4 Inference in the Joint Model

Given the above complex hierarchical model, we now propose a new mean-field based inference approach to perform efficient inference for joint estimation. But, before going into details of our approach, we give a general form of mean-field update. We also highlight the work of Krahenbuhl and Koltun [18] for filter-based efficient inference in fully connected pairwise CRFs. This model was later extended by Vineet et.al. [16] to incorporate higher order potentials, and to solve jointly the object-stereo correspondence problems.

Let us consider a general form of energy function:

$$E(\mathbf{Z}|\mathbf{I}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{z}_c|\mathbf{I}) \tag{10}$$

where $\mathbf{Z}$ is a joint assignment of the random variables $\mathcal{Z} = \{Z_1, ..., Z_{N_Z}\}$, $\mathcal{C}$ is a set of cliques each consisting of a subset of random variables $c \subseteq \mathcal{Z}$, and associated with a potential function ψ_c over settings of the random variables in c, $\mathbf{z}_c$. In Sec. 2 we have that $\mathcal{Z} = \mathcal{X}^S$, that each X_i takes values in the set $\mathcal{L}^S$ of human labels, and that $\mathcal{C}$ contains unary and pairwise cliques of the types discussed. In general, in the models discussed below we will have that $\mathcal{X}^S \subseteq \mathcal{Z}$, so that $\mathcal{Z}$ may also include other random variables (e.g. latent variables) which may take values in different label sets.

Considering this model, the general form of the mean-field update equation (see [15]) is:

$$Q_i(z_i = \nu) = \frac{1}{\tilde{Z}_i} \exp\{-\sum_{c \in \mathcal{C}} \sum_{\{\mathbf{z}_c | z_i = \nu\}} Q_{c-i}(\mathbf{z}_{c-i}) \cdot \psi_c(\mathbf{z}_c)\} \tag{11}$$

where ν is a value in the domain of the random variable z_i, $\mathbf{z}_c$ denotes an assignment of all variables in clique c, $\mathbf{z}_{c-i}$ an assignment of all variables apart from Z_i, and Q_{c-i} denotes the marginal distribution of all variables in c apart from Z_i derived from the joint distribution Q. $\tilde{Z}_i = \sum_\nu \exp\{-\sum_{c \in \mathcal{C}} \sum_{\{\mathbf{z}_c | z_i = \nu\}} Q_{c-i}(\mathbf{z}_{c-i}) \cdot \psi_c(\mathbf{z}_c)\}$ is a normalizing constant for random variable z_i. We note that the summations $\sum_{\{\mathbf{z}_c | z_i = \nu\}} Q_{c-i}(\mathbf{z}_{c-i}) \cdot \psi_c(\mathbf{z}_c)$ in Eq. 11 evaluate the expected value of ψ_c over Q given that Z_i takes the value ν.

Following this general update strategy, the updates for the densely connected pairwise model in Eq. 1 are derived by evaluating Eq. 11 across the unary and pairwise potentials defined in Sec. 2 for $z_i = x_{1...N}$ and $\nu = 0...L$. For the densely connected pairwise CRF model, the mean-field update takes the following form:

$$Q_i(z_i = l) = \frac{1}{\tilde{Z}_i} \exp\{-\psi_i(z_i) - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(z_j = l') \psi_{ij}(z_i, z_j)\} \tag{12}$$

With this mean-field update, Krahenbuhl and Koltun [18] proposed a filter-based method for performing fast inference thus reducing the complexity from $O(N^2)$ to $O(N)$ under the assumption that the pairwise potentials take the form of a linear combination of Gaussian kernels. They show how the expensive message passing update in the mean-field is approximated by a convolution with a bilateral filter in a high dimensional space. Given this Gaussian convolution, they use a permutohedral lattice based bilateral filtering method [20] for performing efficient inference. They run the update equation for a fixed number of iterations, where each iteration leads to a decrease in the KL-divergence value. To extract a solution, they evaluate the approximate maximum posterior marginal as $z_i^* = \max_{z_i} Q_i(z_i)$.

4.1 Efficient Inference

In our framework, we need to jointly estimate the best possible configurations of the segmentation variables X^S, per-pixel part variables X^J, disparity variable X^D and part variable Y^P by minimizing the energy function $E(\mathbf{x}, y)$ in Eq. 5. We now provide the details of our mean-field updates for efficient joint inference.

Update for segment variables (X^S) : Given the energy function detailed in Sec. 3.1, the marginal update for human segmentation variable X_i^S takes the following form:

$$Q_i^S(x_{[i,l]}^S) = \frac{1}{Z_i^S} \exp\{-\psi^S(x_i^S) - \sum_{l' \in \mathcal{L}^J} \sum_{j \neq i} Q_j^S(x_{[j,l']}^S)\psi(x_i^S, x_j^S) \\ - \sum_{l' \in \mathcal{L}^D} Q_i^D(x_{[i,l']}^D)\psi(x_i^D, x_i^S) - \sum_{l' \in \mathcal{L}^J} Q_i^J(x_{[i,l']}^J)\psi(x_i^S, x_i^J)\} \tag{13}$$

where $Q_i^D(x_{[i,l']}^D)\psi(x_i^D, x_i^S)$ and $Q_i^J(x_{[i,l']}^J)\psi(x_i^S, x_i^J)$ are the messages from disparity and per-pixel part variables respectively to the segmentation variables. Thus, these messages enforce the consistency between the segmentation, disparity and per-pixel part term variables. We write $x_{[i,l]}$ for $(x_i = l)$ and the same notation will be followed subsequently.

Update for disparity variables (X^D) : Similar to the updates for X_i^S, the marginal update for the per-pixel depth variables X_i^D takes the following form:

$$Q_i^D(x_{[i,l]}^D) = \frac{1}{Z_i^D} \exp\{-\psi^D(x_i^D) - \sum_{l' \in \mathcal{L}^D} \sum_{j \neq i} Q_j^D(x_{[j,l']}^D)\psi(x_i^D, x_j^D) \\ - \sum_{l' \in \mathcal{L}^S} Q_i^S(x_{[i,l']}^S)\psi(x_i^D, x_i^S) - \sum_{l' \in \mathcal{L}^J} Q_i^J(x_{[i,l']}^J)\psi(x_i^J, x_i^D)\} \tag{14}$$

where $Q_i^S(x_{[i,l']}^S)\psi(x_i^D, x_i^S)$ and $Q_i^J(x_{[i,l']}^J)\psi(x_i^J, x_i^D)$ correspond to the messages from the segmentation and per-pixel part variables to the disparity variables.

Update for per-pixel part variables (X^J) : The per-pixel part variable X_i^J takes messages from part configuration in the hierarchy along with the messages from the other per-pixel part variables, segmentation variables and disparity variables. Thus, the marginal update for the per-pixel part variables X_i^J take the following form:

$$Q_i^J(x_{[i,l]}^J) = \frac{1}{Z_i^J}\exp\{-\psi_u^J(x_i^J) - \sum_{l'\in\mathcal{L}^\mathcal{J}}\sum_{j\neq i} Q_j^J(x_{[j,l']}^J)\psi(x_i^J, x_j^J) \\ - \sum_{l'\in\mathcal{L}^\mathcal{D}} Q_i^D(x_{[i,l']}^D)\psi(x_i^J, x_i^D) - \sum_{l'\in\mathcal{L}^\mathcal{S}} Q_i^S(x_{[i,l']}^S)\psi(x_i^J, x_i^S) - \sum_{l'\in\mathcal{L}^\mathcal{P}} Q_i^P(y_{[i,l']})\psi(y_i, x_i^J)\} \tag{15}$$

Here $Q_i^P(y_{[i,l']})\psi(y_i, x_i^J)$ carry messages from the valid part configuration in the hierarchy to the per-pixel part variables, and $Q_i^S(x_{[i,l']}^S)\psi(x_i^J, x_i^S)$ and $Q_i^D(x_{[i,l']}^D)$ $\psi(x_i^D, x_i^J)$ correspond to the messages from the segmentation and disparity variables to per-pixel part variables.

It is also to be noted that the required expectation update for messages from other joint variables, e.g. messages from segmentation variables to disparity variables, contribute a time complexity of $O(N)$. Thus, the marginal update steps do not increase the overall time complexity.

Update for latent part variables (Y) : Finally, the mean-field update for the part variables in the hierarchy corresponds to:

$$Q_i^P(y_{[i,l]}) \propto \exp\{-\psi_u(y_i) - \sum_{j'\in\mathcal{L}^\mathcal{J}}\sum_{j=1}^{N} Q_j^J(x_{[j,j']}^J)\psi(y_i, x_j^J) \tag{16}$$

where $Q_j^J(x_{[j,j']}^J)\psi(y_i, x_j^J)$ corresponds to the messages from the per-pixel part variables to the part configuration variables. Evaluation of the expectation for part variables contributes $O(N)$ to the overall complexity. Thus, our inference method does not increase the overall complexity of $O(N)$ for fully connected pairwise updates.

5 Learning

The weights $C^{SJ}, C^{SD}, C^{JD}, C^{JP}$ capturing the relationships between variables at different CRF layers are set through cross-validation. Our cross validation step to search for good set of parameters to weight these different terms in Eq. 5 is greedy in the sense that we set them one at a time sequentially. This way of sequential learning ensured an efficient way to search for a good set of the parameters without going through all the possible joint configurations of the parameters. Structured learning [21] provides a possible future direction to learn these parameters, however our focus was efficient inference. Further, the Gaussian kernel parameters are set through cross-validation as well.

6 Experiments

In this section, we demonstrate the efficiency and accuracy provided by our approach on the H2View [1] dataset. Further, to highlight the generalization of our approach, we also conduct experiment on the Buffy [2] dataset where we do not have stereo pairs of images. In all experiments, timings are based on code run on an Intel® Xeon ® 3.33 GHz processor, and we fix the number of full mean-field update iterations to 5 for all models. As a baseline, we compare our approach for the joint estimation of human segmentation, pose, per-pixel part and disparity with the dual-decomposition based model of Sheasby et al. [1]. Further, we compare our joint approach against some other state-of-the-art approaches which do not perform any joint estimation. For example, we compare our human segmentation results against a graph-cuts based AHCRF [22] and the mean-field model of Krähenbühl et al. [18]. We assess human segmentation accuracy in terms of the overall percentage of pixels correctly labelled, the average recall and intersection/union score per class (defined in terms of the true/false positives/negatives for a given class as TP/(TP+FP+FN)). Similarly, for pose estimation, apart from comparing against the dual-decomposition based joint labelling model of Sheasby et.al. [1], we compare the probability of correct pose (PCP) criterion against the models of Yang and Ramanan [7], and Andriluka et al. [23], which do not perform joint labelling. In all these cases, we use the code provided by the authors for the AHCRF, Krähenbühl et al., Yang and Ramanan, Andriluka et al., and Sheasby et al. However we do not quantitatively evaluate the disparity results as we do not have the ground truth data for the disparity.

6.1 H2View dataset

The H2View dataset [1] comprises 1108 training images and 1598 test images consisting of humans in different poses performing standing, walking, crouching, and gesticulating actions in front of a stereo camera. Ground truth human segmentation, and pose are provided; we augment these with a per-pixel part labels.

We first show the accuracy and efficiency achieved by our method on the human segmentation results. We observe an improvement of almost 3.5% over the dual-decomposition based joint inference model of Sheasby et al. [1], almost 4.5% compared to AHCRF [22] and almost 4% over dense CRF [18] in the I/U score, shown in Tab. 1. Significantly, we observe an order of magnitude of speed up (almost 20×) over the model of Sheasby et al. and a speed up of almost 5× over the AHCRF model. Further as far as pose estimation results are concerned, we achieve an improvement of almost 3.5% over Yang and Ramanan, 7% over Andriluka et al. in the PCP scores. Though these methods do not perform joint inference, we compare to highlight the importance of joint inference. Further compared to the model of Sheasby et al., we perform slightly worse in the PCP score, but we observe a speed up of almost 20× over their model. Here it should be noted that the time to evaluate the model of Yang and Ramanan to generate initial pose proposals is not included in the models of Sheasby et.al. and our model. Quantitative results for pose estimation are as shown in Tab. 2. Additionally we observe qualitative improvement in both the segmentation and pose

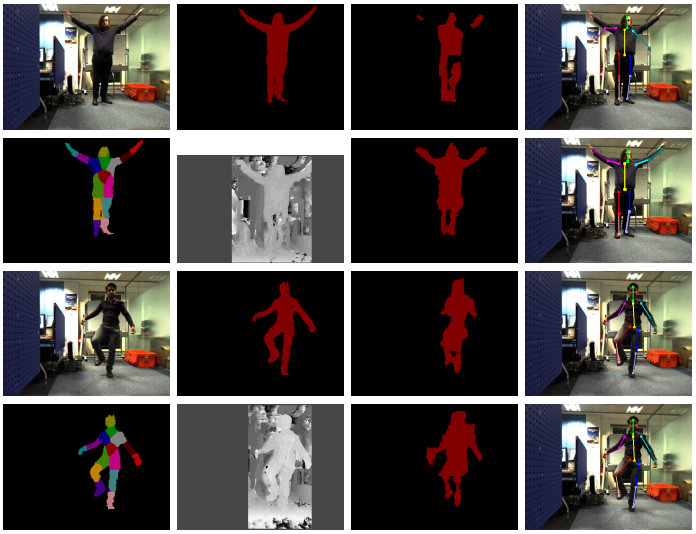

Fig. 3. Qualitative results on two sets of images from H2View dataset. First two rows correspond to the first image, and next two rows to the second image. From left to right: (top row) input image, ground truth for human segmentation, output from [1], pose estimation output from [1]; (second row): our per-pixel part label output, disparity estimation output, segmentation and pose-estimation outputs. Last two rows show the same set of images on the second input image. Our method is able to recover the limbs properly on both the segmentation and pose estimation problems. (Best viewed in color)

results, as shown in Fig. 3. As far as per-pixel part label accuracy is concerned, we achieve 94.43% of overall percentage of correctly labelled pixels, compared to 92.63% achieved by the dual-decomposition method of Sheasby et.al. [1], and 89.55% achieved by the graph-cuts based AHCRF method [22].

Method	Time (s)	Overall	Av.Re	I/U
Unary	0.36	96.12%	85.90%	78.94%
ALE [22]	1.5	96.14%	86.10%	80.14%
Sheasby [1]	25	96.67%	90.48%	81.52%
MF [18]	0.48	96.56%	86.12%	80.57%
Our	1.25	**97.14%**	92.32%	**84.60%**

Table 1. Quantitative results on H2View dataset for human segmentation. The table compares timing and accuracy of our approach (last 2 lines) against the dual-decomposition model of Sheasby et.al. [1] as well as over other baselines. Note the significant improvement in inference time and class-average performance our approach against the baselines.

Fig. 4. Qualitative results on Buffy dataset [2]. From left to right: (first row:) input image, ground truth of segmentation, segmentation output before joint estimation, (second row:) segmentation output after joint estimation, pose output before joint estimation, and pose output after joint estimation. (Best viewed in color)

Method	T(s)	U/LL	U/FA	TO	Head	Overall
Sheasby [1]	25	83.43	54.56	90.05	89.8	73.18
Yang [7]	10	79.65	49.05	88.5	83.0	69.85
Andriluka [23]	35	74.85	47.7	83.9	76.0	66.03
Ours	1.2	82.86	55.16	89.05	86.20	73.12

Table 2. The table compares timing and accuracy of our approach (last line) against the baseline for the pose estimation problem on H2View dataset [1]. Observe that our approach achieves almost 20× speedup against the dual-decomposition model of Sheasby et.al. [1] as well as over other baselines. U/LL represents average of upper and lower legs, and U/FA represents average of upper and fore arms.

6.2 Buffy dataset

In order to show the generalization and effectiveness of our approach, we also evaluate our model on the Buffy dataset. We select a set of 476 images as training images, and 276 images as test images, using the same split as used in [2]. Since there are no depth images, we evaluate only on joint pose and segmentation problems. For the human segmentation case, our joint approach achieves a speed up of almost 70× compared to the dual-decomposition based method of Sheasby et al. [1], and 3× compared to AHCRF [22]. We also observe an improvement of almost 10% and 1% in the I/U scores respectively on segmentation results, shown in Tab. 3. Further, we observe an improvement of almost 0.4% over the Yang and Ramanan model and almost 7% over the model of Sheasby et al. model in the PCP score for the pose estimation problem, shown in Tab. 4. It should be noted that the results of the Yang and Ramanan model [7] reported in our paper is different than the one in their original paper since they first generate a set of detection windows by running an upper-body detector, and then evaluate pose detection only on these detected windows. Here we evaluate the poses on whole image, thus a good detection of the non-detected person could

Method	Time (s)	Overall	Av.Pr	I/U
Sheasby [1]	20	80.85%	85.80%	65.01%
ALE [22]	0.96	87.88%	86.05%	74.16%
MF [18]	0.26	88.40%	**86.47%**	75.01%
Ours	0.28	**88.79%**	86.45%	**75.18%**

Table 3. Quantitative results on Buffy dataset for human segmentation problem. Observe the significant speedup (almost 70×) achieved compared to the dual-decomposition method of Sheasby et.al. [1] and over other approaches. Further, our approach achieves better accuracy than other methods as well.

be penalized. Further improvement through pose estimation within the detected boxes remains a possibility to our approach as well. However, our main goal is to show the efficiency achieved by our joint model without losing any accuracy given the same initial conditions. We also observe an improvement in qualitative results for both the segmentation and pose estimation problems, shown in Fig. 4.

Method	T(s)	L	R	TO	H	Overall
Sheasby [1]	20	61.3	63.5	81.5	85.1	69.17
Yang [7]	1	66.6	71	87.3	90.5	75.6
Ours	0.28	68.2	71	87.6	90.2	76.05

Table 4. The table compares timing and accuracy of our approach (last line) against the baseline for pose estimation problem. Observe that our approach achieves almost 70× speedup, and almost 7% improvement in accuracy over the dual-decomposition model of Sheasby et.al. [1].

7 Discussion

In this work, we proposed *PoseField*, an efficient mean-field based method for joint estimation of human segmentation, pose, per-pixel part and disparity. We formulated this product label space problem in a hierarchical framework, which captures interactions between the pixel level (human/background, disparity, and body part labels), and the part level (head, torso, arm). Finally we have shown the value of our approach on the H2View and Buffy datasets. In each case, we have shown substantial improvement in inference speed (almost $20 - 70\times$) over the current state-of-the-art dual-decomposition methods, while also observing a good improvement in accuracies for both human segmentation and pose estimation problems. We believe our efficient inference algorithm would provide an alternative to some of the existing computationally expensive inference approaches in many other fields of computer vision where joint inference is required. Future directions include investigating new ways to improve the efficiency through parallelization and learning of the relationships between different layers of the hierarchy in a max-margin framework.

Acknowledgment. The work was supported by the EPSRC and the IST programme of the European Community, under the PASCAL2 Network of Excellence. Professor Philip H.S. Torr is in receipt of a Royal Society Wolfson Research Merit Award.

References

1. Sheasby, G., Valentin, J., Crook, N., Torr, P.: A robust stereo prior for human segmentation. ACCV (2012)
2. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008) 1–8
3. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation. In: CVPR, IEEE (2012) 3394–3401
4. Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Brown Univertsity TR **120** (2006)
5. Kumar, M., Zisserman, A., Torr, P.: Efficient discriminative learning of parts-based models. In: CVPR. (2009) 552–559
6. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects (pdf). (2006)
7. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011) 1385–1392
8. Bray, M., Kohli, P., Torr, P.: Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. ECCV (2006) 642–655
9. Kumar, M., Torr, P., Zisserman, A.: Objcut: Efficient segmentation using top-down and bottom-up cues. PAMI **32** (2010) 530–545
10. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Graph cut based inference with co-occurrence statistics. ECCV (2010) 239–253
11. Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.: Joint optimisation for object class segmentation and dense stereo reconstruction. In: BMVC. (2010) 104.1–11 doi:10.5244/C.24.104.
12. Komodakis, N., Paragios, N., Tziritas, G.: Mrf energy minimization and beyond via dual decomposition. (PAMI) 1–1
13. Wang, H., Koller, D.: Multi-level inference by relaxed dual decomposition for human pose segmentation. In: CVPR. (2011) 2433–2440
14. Sheasby, G., Warrell, J., Zhang, Y., Crook, N., Torr, P.: Simultaneous human segmentation, depth and pose estimation via dual decomposition. BMVW (2012)
15. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
16. Vineet, V., Warrell, J., Torr, P.: Filter-based mean-field inference for random fields with higher-order terms and product label-spaces, ECCV (2012)
17. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. (2011) 1297–1304
18. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. (2011) 109–117
19. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV (1). (2006) 1–15
20. Adams, A., Baek, J., Davis, M.A.: Fast high-dimensional filtering using the permutohedral lattice. Comput. Graph. Forum **29** (2010) 753–762
21. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML. (2004)
22. Ladický, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical crfs for object class image segmentation. In: ICCV. (2009) 739–746
23. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. (2009) 1014–1021