Fast Multi-aspect 2D Human Detection

Tai-Peng Tian and Stan Sclaroff[⋆]

Department of Computer Science, Boston University

Abstract. We address the problem of detecting human figures in images, taking into account that the image of the human figure may be taken from a range of viewpoints. We capture the geometric deformations of the 2D human figure using an extension of the Common Factor Model (CFM) of Lan and Huttenlocher. The key contribution of the paper is an improved iterative message passing inference algorithm that runs faster than the original CFM algorithm. This is based on the insight that messages created using the distance transform are shift invariant and therefore messages can be created once and then shifted for subsequent iterations. Since shifting ($O(1)$ complexity) is faster than computing a distance transform ($O(n)$ complexity), a significant speedup is observed in the experiments. We demonstrate the effectiveness of the new model for the human parsing problem using the Iterative Parsing data set and results are competitive with the state of the art detection algorithm of Andriluka, et al.

1   Introduction

We consider the problem of detecting a 2D articulated human figure in a single image. Furthermore, we are interested in recovering the pose of the human figure, where the pose is described by the position and orientation of the legs, arms, torso, and head. This is a difficult problem because the appearance of human figures varies widely due to factors such as clothing, differences in body sizes, articulation of the human body, and viewpoint from which the image is taken. In this paper, we concentrate on modeling the last two factors, i.e., articulation and viewpoint changes.

The prevailing practice is to employ discretization when modeling viewpoint changes and articulations of the human figure. For example, clustering can be used to partition the training data into groups corresponding to different articulation and viewpoint instances [1]. Such an approach is convenient because a simpler single-view or single-configuration model can be used to model the data within each cluster. Unfortunately, there is a price to pay for such a convenience: an additional layer of arbitration logic must be built to coordinate among these models to give an illusion of a multi-aspect and multi-articulation model. This modeling approach is overly complicated and we propose a simpler alternative.

In our approach, we model the geometric deformations of the 2D human figure caused by articulation and viewpoint changes. We separate out these two types of
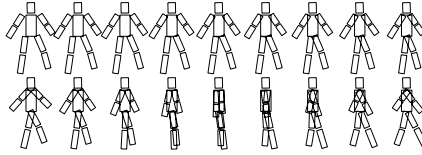

Fig. 1. Fixing the value of the factor for the Common Factor Model (CFM) defines a tree structured Gaussian prior for human poses. Each human pose shown above represents the mean of each distribution for the corresponding value of the factor. In the top row, by varying the factor, the human poses changes from a frontal configuration (leftmost) to a side view (rightmost) configuration. The bottom row depicts the swing of the arms and legs during walking.

deformation into two different modes of variation. These modes can be modeled by a simple extension of the Common Factor Model (CFM) [2] and these modes can be learned using a straightforward training procedure without the need to partition the data into different viewpoints. A concise review of the CFM is given in Sec. 3.

Varying a common factor has the effect of inducing a particular deformation mode in the Pictorial Structure. An intuition for this is given for the human figure model in Fig. 1. If we fix the pose of the human figure and vary the viewpoint by moving along the equator of the view sphere centered on the human figure, then the projected body parts will be translated as the viewpoint changes. Similar observations can be made when a person is walking (viewpoint is kept fixed), which results in rotation and translation of the parts of the Pictorial Structure. This second mode of variation coordinates geometric transformations between body parts; e.g., during a walking cycle the left arm swings forward as the right arm swings backward. Thus, the model of a walking person can be described using a combination of the "walking phase" and "viewpoint" modes. This idea of associating modes of variation with geometric deformations of the Pictorial Structure is general; for example, it is applicable to other types of motion such as a person performing jumping jacks, kicking etc.

Even though CFM inference has linear time complexity, it is still time consuming – especially when the problem size is large, as is the case here. The CFM inference algorithm requires multiple iterations of the min-sum Belief Propagation (BP) algorithm. During each iteration of BP, messages are created from scratch and this is costly because each message contains more than a million entries. Overall, for s iterations of the BP algorithm, there will be $s(n-1)$ messages created for a Pictorial Structure model with n parts.

We propose a new CFM inference algorithm that offers a significant speedup. We reduce the number of messages that need to be created from $s(n-1)$ to $(n-1)$ (a reduction by a factor of s). This speed improvement is significant because the number of BP iterations s scales exponentially in the number of dimensions of the common factor. This speedup relies on two observations: firstly, messages are created using distance transforms and secondly, messages from one iteration

of BP to the next differ only by a shift. Since distance transforms are shift invariant (see proof in Sec 4.1), our method replaces costly distance transforms by shifts, thus gaining a speed improvement over the original formulation. Note that shifting an array only requires an $O(1)$ update to the offset of the array while the distance transform is an $O(h)$ operation that requires visiting all the h elements in the array. Details of the algorithm can be found in Sec. 4.

We provide experimental evaluation of our multi-aspect model in Sec. 6. We show experimental results comparing the speed of our new inference algorithm with the original [2] and evaluate the accuracy of our model on the Iterative Parsing data set [3].

Contribution. The contribution of this paper is twofold. Firstly, we provide a method for modeling multiple modes of deformation in a Pictorial Structure model. Secondly, we improve the running time of the original CFM inference algorithm by observing that messages created by distance transforms are shift invariant. Replacing costly $O(h)$ time complexity distance transforms with fast $O(1)$ time complexity shifting yields a significant speed up.

2   Related Work

Our work is related to the use of Pictorial Structures for detecting human figures using tree structured Gaussian graphical models [3–6], as well as loopy graphical model [7–10]. Our work is different from these related work as we focus on modeling geometric deformation of the Pictorial Structures due to factors such as viewpoint changes and phase of the walking cycle.

Our work builds on the Common Factor Model (CFM) [2]. Originally in [2], a 1D latent variable (or factor) is used to model the phase of a walking cycle, and it is used to capture correlations among the upper limbs of the human figure. We provide a new perspective on the CFM by interpreting the dimensions of the factor as modes of geometric deformation in the Pictorial Structure.

Unfortunately, using higher dimensional latent variables increases the CFM inference running time, e.g., if uniformly sampling the 1D factor requires n samples then in 2D it will require n^2 samples. This slows down the CFM inference significantly because multiple distance transforms are required in each iteration of the inference algorithm. We propose a faster CFM inference that only requires a constant number of distance transforms to be computed, i.e., independent of the number of iterations in the CFM inference.

Other multi-aspect modeling works [1, 11, 12] use a discrete set of viewpoints. In contrast, our work uses a continuously parameterized viewpoint.

3   Background: The Common Factor Model

In this section, we review the Common Factor Model (CFM) of [2]. The CFM provides an alternative to high order clique models. Such high order clique models arise in 2D human detection because strong correlations exist among the
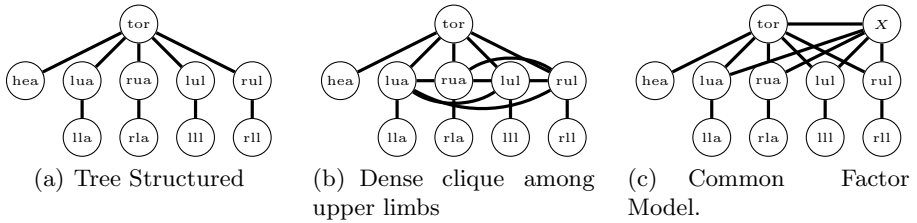
(a) Tree Structured   (b) Dense clique among upper limbs   (c) Common Factor Model.

Fig. 2. Different type of priors used for the ten part human figure. Abbreviations are *tor* : torso, *hea*: head, *lua* : left upper arm, *rll* : right lower leg, etc.

upper arms and upper legs when a person is walking [2]. These dependencies create a large clique among the upper limbs of the graphical model (Fig. 2(b)) and inference over graphical models with large cliques is computationally intensive. The computational difficulty can be ameliorated by breaking up the large clique into smaller cliques. This breaking up is justified by positing that a latent variable X is responsible for the observed correlations among the upper limbs (Fig. 2(c)). More importantly, when the latent variable X is observed, i.e., conditioned on X, then the graphical model becomes a tree again. The latent variable X can be viewed as a hyper parameter and fixing a value for this hyper parameter will produce the tree structured Gaussian prior in Fig. 2(a), but parameters for this tree structured prior will be different for two distinct values of X.

The detection problem is stated as finding the latent variable value X^* and body parts locations L^* that maximize the posterior, i.e.,

$$\langle L^*, X^* \rangle = \arg\max_{L,X} p(L, X | I) = \arg\max_{L,X} p(I | L, X)\, p(L, X), \qquad (1)$$

where I is the image, $L = \{l_i\}$ and i are body part names corresponding to nodes shown in Fig 2(a). Each body part configuration l_i is described by an oriented rectangle comprising the center of the rectangle (u, v) and its orientation θ.

The CFM takes on the following factorization

$$p(I|L, X)p(L, X) = p(I|L, X)p(L|X)p(X)$$

$$\propto \underbrace{\prod_{i \in V} p(I|l_i)}_{\text{likelihood}} \left(\prod_{e_{ij} \in E_X} \phi_{ij}(l_i, l_j, X) \underbrace{\prod_{e_{ij} \in E_T - E_X} \phi_{ij}(l_i, l_j)}_{\text{prior}} \right) p(X),$$

where the likelihood is independent of the latent variable X and the CFM assumes that image appearances among body parts l_i are independent. In the above equation, V is an index set for the body parts of the 2D human model, which corresponds to the set of vertices shown in Fig. 2(a). The set of edges E_T is shown in Fig. 2(a), and E_X is a subset of E_T. Edges from E_X have both end vertices forming a clique with the latent variable X in Fig. 2(c). The prior is factorized according to the graphical model in Fig 2(c), and parameters for the common factor X are learned from data [2]. The compatibility function ϕ_{ij}

between two body parts is defined based on the distance Y_{ij} between the joint locations $T_{ij}(l_i)$ and $T_{ji}(l_j)$, i.e.,

$$Y_{ij} = T_{ij}(l_i) - T_{ji}(l_j). \tag{2}$$

The transformation T_{ij} shifts the body part center to the joint position, i.e.,

$$T_{ij}(l_i) = T_{ij}([u, v, \theta]^T) = [u', v', \theta]^T, \qquad \text{where} \quad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} + R_\theta \begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix}. \tag{3}$$

In the above equation, R_θ is the rotation matrix by θ angle, u_{ij} and v_{ij} are connection parameters that are learned from a tree structured prior [5]. The definition for the transformation T_{ji} is similar to T_{ij} and details are given in [5]. For edges that are not involved with the common factor, the compatibility function is given by

$$\phi_{ij}(l_i, l_j) = N\left(Y_{ij}; 0, \Sigma_{ij}\right), \tag{4}$$

where Σ_{ij} is a diagonal matrix learned from data [5], and N is the Gaussian function. For edges that are involved with the common factor X, the potential function is given as

$$\phi_{ij}(l_i, l_j, X) = N\left(Y_{ij} - A_j X; 0, \Psi_j\right), \tag{5}$$

where A_j is part of the loading matrix A learned from data. Both of these are defined in the next paragraph.

Learning the Loading Matrix A: In order to learn the loading matrix A, the training data for the four body parts l_{lua}, l_{rua}, l_{lul}, l_{rul} are stacked up into a 12 dimensional vector. Suppose there are m training instances, then a $12 \times m$ matrix is formed and Common Factor Analysis is applied on this matrix to recover the loading matrix A and covariance matrix Ψ. If the dimension of the common factor X is two, then the resulting loading matrix A will have dimension 12×2, and the covariance matrix Ψ will be a 12×12 matrix. Therefore, A_{lul} denotes the corresponding 3×2 sub matrix of A whose rows correspond to the stacking order for the body part left upper leg (lul). The covariance sub-matrix Ψ_{lul} will be a 3×3 square matrix that includes the diagonal entries of Ψ whose rows correspond to the stacking order for lul.

3.1   Messages and Dynamic Programming in the CFM

In this section, we review the message passing algorithm applied on the tree structured model generated by the Common Factor Model (CFM). In the CFM inference, the goal is to find the best body part location L^* and common factor X^* that maximize the posterior $p(L, X|I)$. This is equivalent to minimizing the negative log posterior, which is

$$\langle L^*, X^* \rangle = \arg\min_{L,X} c(X) + \sum_{i \in V} m_i(l_i) + \sum_{ij \in E_T - E_X} d_{ij}(l_i, l_j) + \sum_{ij \in E_X} d_{ij}(l_i, l_j, X), \tag{6}$$
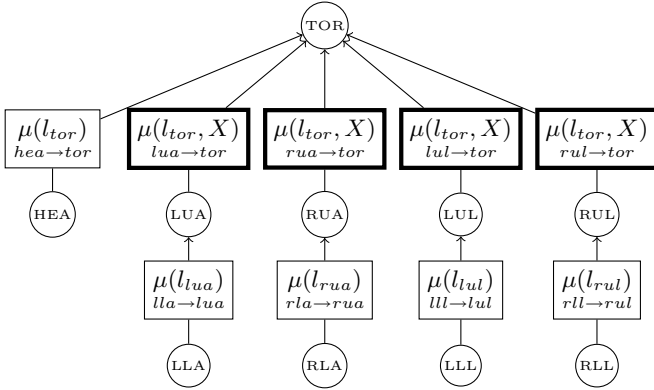
Fig. 3. The boxes show the messages passed during an iteration of the Belief Propagation algorithm for a fixed value of the common factor X. Bold boxes indicate messages parameterized by the common factor X.

where $c(\cdot)$ is the negative log of the prior $p(X)$, $m_i(\cdot)$ is the negative log of the likelihood, and $d_{ij}(\cdot)$ is the negative log of the compatibility function.

Given a fixed value X for the common factor, the resulting graphical model is a tree. Therefore, dynamic programming can be used to find the MAP solution. The dynamic programming proceeds from leaves to the root and intermediate results from the dynamic programming can be interpreted as messages being passed from leaves up to the root. These messages can be efficiently computed via the distance transform [5]. The types of messages passed between a child node j and its parent parent i are

$$
\underset{j \to i}{\mu}(l_i) = \begin{cases} \mathcal{T}_{ij}^{-1} \mathcal{D} \mathcal{T}_{ji}[m_j](l_i), & \text{if } j \text{ is a leaf node,} \\ \mathcal{T}_{ij}^{-1} \mathcal{D} \mathcal{T}_{ji}\left[m_j + \sum_{c \in C_j} \underset{c \to j}{\mu}\right](l_i), & \text{if } j \text{ is an internal node with children } C_j, \\ \mathcal{T}_{ij}^{-1} \mathcal{D} \mathcal{T}_{A_j X} \mathcal{T}_{ji}\left[m_j + \sum_{c \in C_j} \underset{c \to j}{\mu}\right](l_i), & \text{if } j \text{ is an internal node with children } C_j \\ & \quad \text{and common factor } X, \end{cases}
\tag{7}
$$

where $\mathcal{T}_{ij}$ and $\mathcal{T}_{ji}$ are operators that bring the coordinates of body parts into ideal alignment at the joint, $\mathcal{T}_{A_j X}$ is the translation induced by the common factor X, and $\mathcal{D}$ is the distance transform operator. All of these are defined as

$$
\mathcal{T}_{ij}^{-1}[f](l_j) = f(T_{ij}^{-1}(l_j)), \qquad \mathcal{T}_{ji}[f](l_j) = f(T_{ji}(l_i)), \tag{8}
$$
$$
\mathcal{T}_{x_j}[f](l_j) = f(l_j - A_j X_j), \qquad \mathcal{D}[f](l_j) = \min_{l_i \in \mathcal{G}} f(l_i) + ||l_i - l_j||^2,
$$

where $\mathcal{G}$ represents grid positions on which the function f is sampled. Note the notational difference between $\mathcal{T}_{ij}$ (in calligraphic script) and T_{ij} (in regular font); they are conceptually different as the operator $\mathcal{T}_{ij}$ transforms one function into another, whereas the function T_{ij} transforms coordinates. Lastly, the operators

are applied from right to left, i.e., for the chain of operations, $\mathcal{T}_{ij}^{-1} \mathcal{T}_{A_j X} D \mathcal{T}_{ji}[f]$, the operator $\mathcal{T}_{ji}$ is applied first, followed by D, $\mathcal{T}_{A_j X}$ and $\mathcal{T}_{ij}^{-1}$.

These messages are depicted in Fig. 3. At the root node, the messages are combined, and the best configuration for the root is

$$l_{tor}^* = \min_{l_{tor}} \left(m_{tor}(l_{tor}) + \sum_{c \in C_{tor}} \mu_{c \to tor}(l_{tor}) \right). \tag{9}$$

Once the best solution for the root is found, the algorithm backtracks down the tree to recover the corresponding values for other body parts.

4 Faster Inference for the Common Factor Model

We propose a method that speeds up the inference algorithm of Lan and Hut-tenlocher [2]. First we briefly review the inference algorithm of Lan and Hut-tenlocher. During inference, values are sampled from the latent variable X and for each sample value, an iteration of dynamic programming (DP) is performed. For each DP iteration, the messages are created from scratch by applying distance transforms [5]. Overall, the number of distance transforms required scales linearly with the sample size of the common factor, i.e., $s(n-1)$ distance transforms are required, where s is the sample size for the common factor X and n is the number of body parts.

We propose a method that reduces the number of distance transforms required. Our method only requires computing $n - 1$ distance transforms, i.e., independent of the number of samples size s for X. This is a significant speedup because s scales exponentially in the dimension of the of the common factor X. This speedup is possible because varying the values of X has the effect of shifting the messages, and secondly, distance transforms are shift invariant. Therefore, new messages can be created by simply shifting the previous messages. Computationally, shifting is more efficient than DT because shifting has $O(1)$ time complexity (where we only need to update the offset for the array), compared to $O(h)$ time complexity for DT, where the algorithm has to visit all the h entries in the array (typically, $h \sim 10^6$ for the examples we are testing on).

The next section gives the proof for shift invariance of distance transforms and following that, we describe the inference algorithm.

4.1 Distance Transforms Are Shift Invariant

We prove that the distance transform of a sampled function is shift-invariant under some fairly mild conditions that are usually satisfied in practice. Let $\mathcal{D}$ be the distance transform operator, where

$$\mathcal{D}[f](p) = \min_{q \in \mathcal{G}} f(q) + ||p - q||^2, \tag{10}$$

and p is a position in the grid $\mathcal{G}$ for sampling the function f. The operator $\mathcal{T}_r$ translates a function f by r, that is,

$$\mathcal{T}_r[f](p) = f(p + r). \tag{11}$$

Proposition 1. *Suppose f is a function sampled on the grid $\mathcal{G}$. For any given position $p \in \mathcal{G}$ and a fixed translation r, such that $f(p) = \infty$ if $p \notin \mathcal{G}$, then $\mathcal{DT}_r[f](p) = \mathcal{T}_r\mathcal{D}[f](p)$.*

Proof. Starting from LHS,

$$\mathcal{DT}_r[f](p) = \mathcal{D}[g](p) \qquad \text{(where } g(x) \equiv f(x + r)) \tag{12}$$

$$= \min_{v \in \mathcal{G}} g(v) + ||p - v||^2 \tag{13}$$

$$= \min_{v \in \mathcal{G}} f(v + r) + ||p - v||^2 \tag{14}$$

$$= \min_{(q-r) \in \mathcal{G}} f(q) + ||p + r - q||^2. \qquad (q = v + r) \tag{15}$$

On the RHS,

$$\mathcal{T}_r\mathcal{D}[f](p) = \mathcal{T}_r[h](p) \qquad \text{where } h(p) \equiv \min_{h \in \mathcal{G}} f(q) + ||p - q||^2 \tag{16}$$

$$= h(p + r) \tag{17}$$

$$= \min_{q \in \mathcal{G}} f(q) + ||p + r - q||^2. \tag{18}$$

Therefore, the operator $\mathcal{D}$ commutes with the operator $\mathcal{T}_r$. □

4.2 Faster Inference

We describe how to exploit the shift invariance property of the distance transform to speed up the inference algorithm. Within different iterations of the inference algorithm, messages originating from the leaves do not change (Fig. 3); only messages affected by the common factor X are recomputed. Those messages affected by the common factor are recomputed using the chain of operators $\mathcal{T}_{ij}^{-1}\mathcal{DT}_{x_j}\mathcal{T}_{ji}$. Notice that the distance transform operator $\mathcal{D}$ is applied *after* the translation operator $\mathcal{T}_{x_j}$; therefore, based on this chain of operations, when the common factor X changes, a distance transform operation is required to compute the new message. Since the distance transform is shift invariant, we can rewrite the messages involving the common factor X as

$$\mu_{j \to i}(l_i) = \mathcal{T}_{ij}^{-1}\mathcal{T}_{x_j}\mathcal{DT}_{ji}[f](l_i), \qquad \text{where } f = m_j + \sum_{c \in C_j} \mu_{c \to j}, \tag{19}$$

where the positions of the operators $\mathcal{D}$ and $\mathcal{T}_{x_j}$ are swapped, i.e., the operator $\mathcal{D}$ has been pushed inwards to the right. Conceptually, this means that we can memoize the result of $\mathcal{DT}_{ji}[f]$ as this does not vary with the common factor X, and for varying X, we only need to apply the operator $\mathcal{T}_{ij}^{-1}\mathcal{T}_{x_j}$ to the memoized
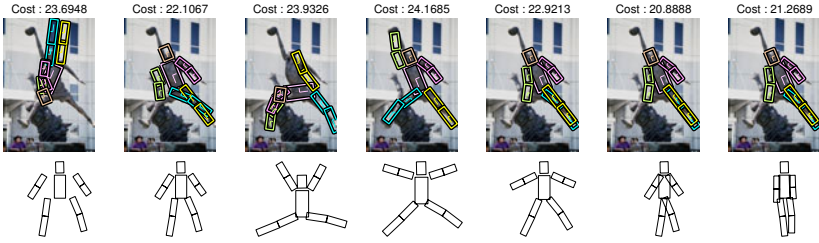
Fig. 4. Human pose prior affects the detection results. **Row 1** shows the optimal pose detected. **Row 2** shows the mean of the tree structured Gaussian prior for the human pose. Notice that the most visually appealing solution (center image) does not correspond to the configuration with the lowest cost.

$\mathcal{DT}_{ji}[f]$. Computationally, this translates to substantial savings because for each new message to be created, we only require the translation operator $\mathcal{T}_{ij}^{-1}\mathcal{T}_{x_j}$. Overall, only $n-1$ distance transformed messages need to be computed, for n body parts, compared to $s(n-1)$ originally, where s is the number of samples for the common factor X.

5   Detection Using the Multi-aspect Model

Computing the the maximum *a posteriori* (MAP) estimate or, equivalently, finding the lowest cost configuration does not necessarily give the most visually correct solution (see the example in Fig. 4). We remedy this problem using a "sample and test" strategy [5, 13]. First, we sample a set of values for the factors of the CFM and recover the corresponding set of detection results. Following that, detection results are re-evaluated using additional constraints. We summarize the detection algorithm in Algorithm. 1. The following constraints are used to re-score the detection results.

1. **Appearance Symmetry:** Humans typically wear clothing that is symmetric and we penalize detection results with dissimilar appearance between the upper arms and upper legs of the Pictorial Structure. Dissimilarity of appearance between two body parts is described using the distance between the two Region Covariance (RC) descriptors [14]. The RC descriptor for a body part is a 5×5 symmetric matrix and involves entries for spatial positions (x, y) and the three color channels of the image (r, g, b). The distance ρ_1 between two RC descriptors C_1 and C_2 is given as

$$\rho_1(C_1, C_2) = \gamma \sqrt{\sum_{i=1}^{5} \lambda_i(C_1, C_2)}, \qquad (20)$$

where $\{\lambda_i(C_1, C_2)\}_{i=1\cdots5}$ are the generalized eigenvalues of C_1 and C_2, and γ is a scaling factor chosen empirically to be 0.1.

Algorithm 1. Detection Algorithm for the Multi-Aspect Model

Let $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$ be the samples for the common factor.

Let $C = \{lua, rua, lul, rul\}$.

Let $pairs = \{(lua, rua), (lul, rul)\}$

Compute the messages $\underset{j \to i}{\mu}$ shown in Fig. 3 with $X = 0$.

for $k = 1 \ldots s$ **do**

$$\underset{tor}{\mu'}(p) = \underset{hea \to tor}{\mu}(p) + \sum_{i \in C} T_{\mathbf{A_i X_i}} \left[\underset{i \to tor}{\mu} \right](p) \qquad (T_\mathbf{r}[\cdot] \text{ in Eqn. 11 and } A_i X_i \text{ in Eqn. 5})$$

$$p^* = \arg\min_{tor} \mu'(p)$$

$$score(k) = \underset{tor}{\mu'}(p^*) + \sum_{ij \in pairs} \rho_1(l_i, l_j) + \sum_{ij \in pairs} \rho_2(l_i, l_j) \qquad (\rho_1, \rho_2 \text{ Eqn. 20,21})$$

end for

$bestscore = \min score(k)$

To recover the pose with the best score, perform a backtracking on the corresponding messages (similar to backtracking for dynamic programming [5]).

---

2. **Overlapping Bodyparts:** Tree structured Pictorial Structures are prone to the "over counting of evidence" problem, e.g., the legs typically snap onto the same region in the image. We can ameliorate this problem by adding a penalty term

$$\rho_2(l_i, l_j) = \frac{|R(l_i) \bigcap R(l_j)|}{\min(|R(l_i)|, |R(l_j)|)} \tag{21}$$

for overlapping body parts, where l_i and l_j are the configurations of body parts i and j, $R(\cdot)$ denotes the rectangular region in the image covered by the configuration of a body part and $|\cdot|$ denotes the area. The overlap area is computed by first clipping the rectangle $R(l_i)$ against $R(l_j)$ using the Sutherland Hodgman clipping algorithm and the resulting polygon gives the overlapping region. The overlap area is scaled to the range $[0, 1]$ by dividing it by the smaller body part's area.

6   Experiments

We use the Iterative Parsing (IP) data set [3] for all the experiments. This challenging data set contains a large variety of human figures in difficult poses such as baseball pitchers, sumo wrestlers, etc. The Pictorial Structure parameters are learned from data following [5]. For the body parts detector, we use the code from [4]. All coding is done in Matlab and the computationally intensive functions such as distance transforms are implemented in mex code.

For the common factor, we learned a two-dimensional common factor from the training set in the IP data set. We were able to obtain the viewpoint effect, i.e., varying the first common factor adjusts the joint position between the upper arms / legs to be closer or further apart, giving the effect of a viewpoint change from side view to front view (see Fig. 1). Unfortunately, the training data does not contain sufficient variations in the swing of the arms and legs to learn a
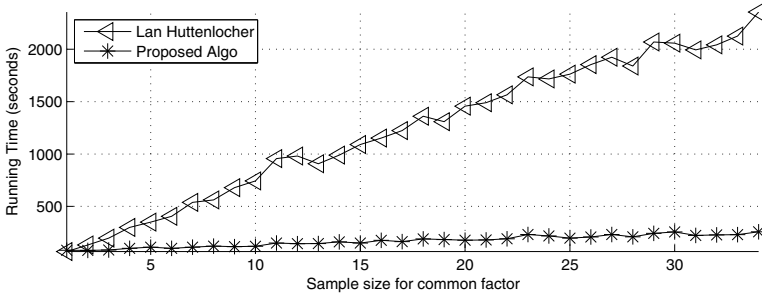
Fig. 5. Comparing the running time for Lan and Huttenlocher's [2] inference algorithm with the proposed algorithm for various sample sizes for the common factor. Both algorithms have linear running time curves but the proposed algorithm is faster, e.g., six times speedup for 10 samples, eight times speedup for 20 samples and nine times speedup for 35 samples. The speedup continues to grow for increasing sample sizes.

common factor for that effect; in contrast, [2] uses primarily walking sequences as training data and is able to capture the arm swing effect in the common factor. As a substitute, the following loading matrix is used in all the the experiments,

$$A = \begin{bmatrix} A \\ tor,lua \end{bmatrix}, \underset{tor,rua}{A} \underset{tor,lul}{A}, \underset{tor,rul}{A} \end{bmatrix}^T, \tag{22}$$

where

$$\underset{tor,lua}{A} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \underset{tor,rua}{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \underset{tor,lul}{A} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \underset{tor,rul}{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -0.5 \end{bmatrix}.$$

For each sub matrix, the three columns are ordered according to (u, v, θ), where (u, v) is the spatial location and θ is the rotation angle. The loading matrices above can be considered as idealized versions of those learned from the IP data set, as well as the the loading matrix published in [2].

Speed Comparison. We compare the running time of the proposed algorithm against [2]. We fix the image (size 454×353) and vary the number of samples for the common factor. The plot of running times versus varying samples for the common factor is shown in Fig. 5. Asymptotically, both algorithms have linear time complexity, but empirically, the proposed algorithm runs significantly faster in practice. For example, when using 10 samples, we observe a six fold speedup (120 seconds vs. 743 seconds). The speed gap between the two algorithms continues to widen as the number of samples is increased, e.g., at 20 samples we observe an eight-fold speedup, and at 35 samples there is a nine-fold speedup. This linear increase in speedup trend is true for increasing number of samples.

Accuracy of Parts Localization. We compare the accuracy of localizing body parts for our algorithm against three state of the art algorithms: the standard PS model [5], the Common Factor Model [2] and the work of Andriluka, et al. [4].

Table 1. Body part detection accuracy in percentages. A body part is correctly localized when both ends of the limb are within half the part's length from the ground truth (similar to [4, 6]). **(Row 1)** The standard pictorial structures model with a tree structured prior. **(Row 2)** The Common Factor Model. **(Row 3)** Our proposed multi-aspect detection that includes appearance symmetry and rectangle overlap constraints. **(Row 4)** Andriluka (AN), et al. [4]. The results obtained for AN differ slightly from published result because we used our own implementation of the algorithm.

	Torso	Upper Arms		Upper Legs		Lower Arms		Lower Legs		Head	Avg
		Left	Right	Left	Right	Left	Right	Left	Right		
FH [5]	67.8	32.7	35.1	58.0	52.7	24.4	27.3	54.6	42.9	37.6	43.3
CFM [2]	78.5	41.0	42.0	63.9	59.5	30.2	28.3	62.4	46.8	53.7	50.6
Our work	**80.0**	41.0	41.0	**65.9**	**62.4**	31.2	30.0	**62.4**	**47.8**	54.1	51.6
AN [4]	79.0	**45.9**	**47.8**	65.4	59.0	**33.7**	**34.1**	61.4	47.3	**57.6**	**53.1**

In the experiments, the Common Factor Model and our multi-aspect model use the same parameter for the prior. Samples are drawn from the 2D common factor X as follows. First, we sample the first dimension (controlling the aspect) while keeping the other dimension fixed and values are sampled in the range $[-22, 15]$ at increments of 1.5 resulting in 26 samples. Next, we sample the other dimension that coordinates the swinging of the arms and legs while keeping the first dimension fixed. Values are sampled in the range $[\frac{-18\pi}{17}, \frac{13\pi}{17}]$ in increments of $\frac{\pi}{17}$ resulting in 26 samples. Overall, there are 52 samples chosen for the common factor X. We have found that uniformly sampling the 2D grid to generate 26^2 samples is excessive for the walking human figure model; e.g., from a front view, deformation of the Pictorial Structure due to walking is small. In contrast, these deformations are more prominent from a side view. Therefore, we concentrate on capturing prominent deformations in our sampling.

The Common Factor Model picks the maximum *a posteriori* solution over these 52 samples, but our multi-aspect model re-scores the solution using the ρ_1 and ρ_2 (Sec. 5), and picks the solution with the minimum cost. The localization results are summarized in Table 1. A part is classified as correctly localized when both endpoints of that body part are within 50% of the length of the body part (similar to [4, 6]).

Our approach (Row 3, Table 1) yields better localization results when compared with the standard Pictorial Structures (Row 1 FH) for all the body parts. The best improvement is in the localization of the left upper leg, which shows an increase in correct detections of 13.9%. This is because the standard Pictorial Structure uses a tree structured Gaussian prior that is biased towards a frontal view, and it is prone to the "over counting of evidence" problem.

When compared against the Common Factor Model (Row 2, Table 1), our results (Row 3, Table 1) show an improvement in correct detection that averages about 2% across all the body parts. The difference between the two algorithms is in the inference step. CFM uses the MAP solution, but we re-score the solutions using additional constraints therefore improvements in the detection results are attributed to the re-scoring step. Qualitative examples are shown in Fig. 6.
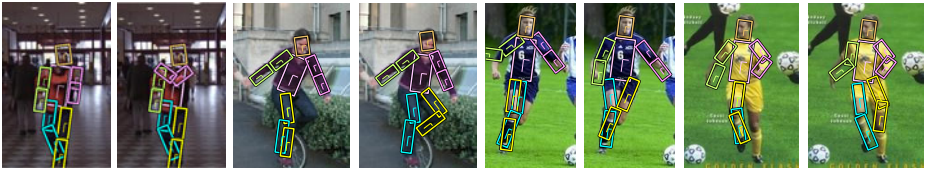
Fig. 6. Examples where incorporating appearance symmetry and rectangle overlap constraints improve detection results. In each pair of image the **left image** shows the detection result using the Common Factor Model [2] and the **right image** shows the detection result obtained using our multi-aspect model. For example, in the first pair of images, the person's left arm is across the chest and this is correctly detected by our method.
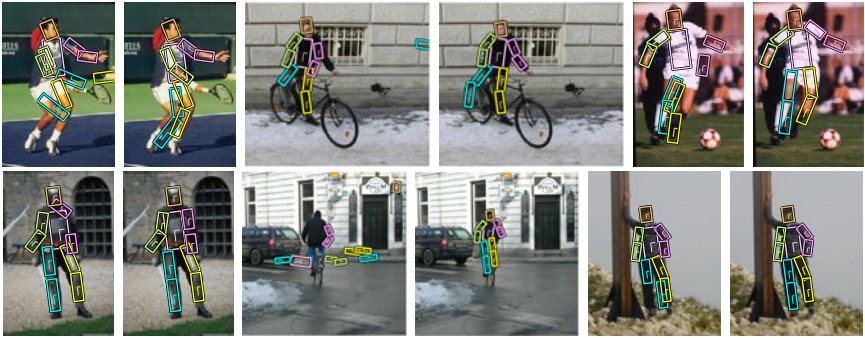

Fig. 7. Examples of the "scattered body parts" problem present in Andriluka, et al.'s [4] detection method. In each pair of image the **left image** shows the detection result using Andriluka, et al.'s method and the **right image** shows the detection result obtained using our multi-aspect model.

We have mixed results when comparing with Andriluka (AN), et al. [4] (Row 4, Table 1). AN has better results for localizing upper and lower arms while we have better results for localizing upper and lower legs. We found that AN's approach suffers from the "scattered body parts" problem, which arises because AN's inference algorithm maximizes the marginal posterior and spatial constraints between body parts are not strictly enforced. This results in solutions where body parts are not tightly grouped together. We show more of these examples in Fig. 7. Our detection results do not suffer from this problem.

7   Conclusion

We have presented a multi-aspect model that is capable of capturing the effects of viewpoint changes in Pictorial Structures using an extension of the Common Factor Model (CFM). We also proposed a two stage algorithm that rescores CFM solutions using additional constraints and this method is shown to be effective in the experiments. Furthermore, we demonstrate how to exploit the

shift invariance property of distance transforms to provide a speedup for the CFM inference algorithm; consequently, we can sample a larger set of samples for the common factor during CFM inference. Sampling a larger set of samples for the common factor enables testing of more views during inference, which contributes to the improved detection results in our experiments.

References

1. Seemann, E., Leibe, B., Schiele, B.: Multi-aspect detection of articulated objects. In: Proc. CVPR (2006)
2. Lan, X., Huttenlocher, D.P.: Beyond trees: Common-factor models for 2D human pose recovery. In: Proc. ICCV (2005)
3. Ramanan, D.: Learning to parse images of articulated objects. In: Proc. NIPS (2006)
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Proc. CVPR (2009)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV 61, 55–79 (2005)
6. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Proc. CVPR (2008)
7. Ren, X., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: Proc. ICCV (2005)
8. Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. In: Proc. CVPR (2008)
9. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: Proc. CVPR (2005)
10. Bergtholdt, M., Kappes, J., Schmidt, S., Schnorr, C.: A study of part-based object class detection using complete graphs. IJCV 28, 416–431 (2009)
11. Kumar, M.P., Koller, D.: Learning a small mixture of trees. In: Proc. NIPS (2009)
12. Lan, X., Huttenlocher, D.: A unified spatio-temporal articulated model for tracking. In: Proc. CVPR (2004)
13. Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Long term arm and hand tracking for continuous sign language TV broadcasts. In: Proc. BMVC (2008)
14. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006)