Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation

Leonid Sigal Michael J. Black
Department of Computer Science, Brown University, Providence, RI 02912
{ls,black}@cs.brown.edu

Abstract

Part-based tree-structured models have been widely used for 2D articulated human pose-estimation. These approaches admit efficient inference algorithms while capturing the important kinematic constraints of the human body as a graphical model. These methods often fail however when multiple body parts fit the same image region resulting in global pose estimates that poorly explain the overall image evidence. Attempts to solve this problem have focused on the use of strong prior models that are limited to learned activities such as walking. We argue that the problem actually lies with the image observations and not with the prior. In particular, image evidence for each body part is estimated independently of other parts without regard to self-occlusion. To address this we introduce occlusion-sensitive local likelihoods that approximate the global image likelihood using per-pixel hidden binary variables that encode the occlusion relationships between parts. This occlusion reasoning introduces interactions between non-adjacent body parts creating loops in the underlying graphical model. We deal with this using an extension of an approximate belief propagation algorithm (PAMPAS). The algorithm recovers the real-valued 2D pose of the body in the presence of occlusions, does not require strong priors over body pose and does a quantitatively better job of explaining image evidence than previous methods.

1. Introduction

Recent approaches to articulated human body detection and pose estimation exploit part-based tree-structured models [3, 5, 8, 13, 15, 17] that capture kinematic relations between body parts. In such models a body part is represented as a node in a graph and edges between nodes represent the kinematic constraints between connected parts. These models are attractive because they allow local estimates of limb pose to be combined into globally consistent body poses. While this distributed computation admits efficient inference methods, the local nature of the inference itself is also the Achilles heal of these methods. The image evidence

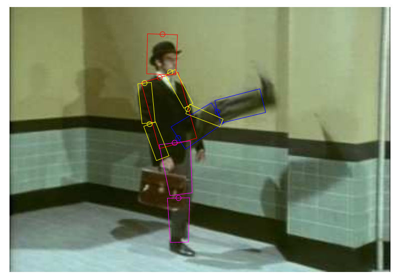


Figure 1. **Silly Walks.** The detection of 2D body pose in real images is challenging due to complex background appearance, loose monochromatic clothing, and the sometimes unexpected nature of human motion. In this scene, strong, activity-dependent, prior models of human pose are too restrictive. The result here was found by our method which makes weak assumptions about body pose but uses a new occlusion-sensitive image likelihood.

for each part is estimated independently of the other parts and, without a global measure of the image likelihood of a body pose, multiple body parts can, and often do, explain the same image data.

In particular, for 2D body pose estimation, the "wrong" solutions are often more likely than the "true" solution. Figure 2 illustrates the problem that results when local image likelihood measures for each body part do not take into account the poses of other parts and do not exploit any knowledge of what image evidence is left unexplained. This problem is not unique to human pose estimation and applies in other generic object-recognition problems.

Recent attempts to solve the problems illustrated in Figure 2 have focused on the use of strong prior models of body pose that rule out unlikely poses [8]. These approaches are not appropriate for dealing with unexpected or unusual motions such as those in Figure 1. In particular, they require that we already know the activity being observed and that the variation in the pose is within learned limits. Other computational strategies incrementally explore the space of body poses but give up the formal probabilistic interpretation of graphical models [13]. In this paper we argue that such approaches are fighting the wrong image likelihood

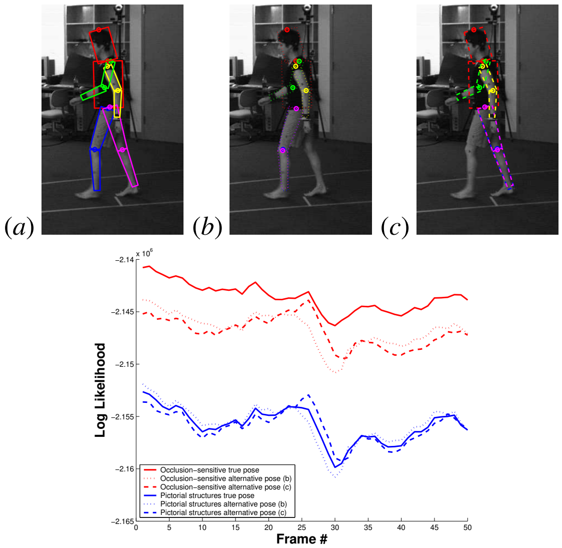


Figure 2. **Fighting the Likelihood**. (*a*) shows the ground truth body pose while (*b*) and (*c*) show common failure modes of pictorial structure approaches in which both legs explain the same image data. With local image likelihoods, the poses in (*b*) and (*c*) are often better interpretations of the scene than the true pose. This can be seen in the plot where 50 frames of a test sequence are evaluated. The blue curves illustrate the local pictorial structures likelihood. The likelihood of the ground truth is solid blue while the likelihoods for the two alternative poses (both legs front or both legs back) are shown as dashed lines. The local likelihood marginally prefers the true pose in only 2 out of 50 frames tested. With our proposed occlusion-sensitive likelihood (shown in red) the true pose is always more likely than the alternative poses.

and that the solution lies in the proper formulation of this likelihood function. A fully global likelihood is computationally impractical and consequently we develop a principled approximation to the global likelihood that is sensitive to local occlusion relationships between parts.

Our contribution is two fold: (1) we introduce a general framework for developing occlusion-sensitive likelihoods that attempt to explain as much of the image as possible, and (2) since occlusion reasoning involves interactions between non-adjacent body parts which create loops in the graphical model structure representing the body, we introduce a variant of approximate belief propagation (BP) that is able to infer the real-valued pose of the person in 2D.

2. Related Work

Generative, model-based, approaches for recovering 2D articulated pose can be loosely classified into two categories. Top-to-bottom approaches treat the body as a "cardboard person" [7] in which the limbs are represented by 2D patches connected by joints. These patches are connected in a kinematic tree [1, 2, 4, 11, 14, 18, 20] and the pose of the person is represented by a high-dimensional state vector that includes the position and orientation of the root limb in the global image coordinate frame and the parameters of each limb relative to its parent in the tree. The high-dimensional state space makes exhaustive search for the body pose difficult. Lee and Cohen [9] address this using a bottom-up proposal process and inverse kinematics to explore the parameter space in a data-driven MCMC procedure. Currently their methods appear to be limited to frontal poses where most body parts are unoccluded.

In contrast, bottom-up approaches address the dimensionality of the state space by representing each part independently in the 2D image coordinate frame. In such models a body part is represented as a node in a graph and edges in the graph represent kinematic constraints between connected parts. This formulation allows independent search for the parts which are then combined subject to the kinematic constraints. The results are typically imprecise, but enable automatic initialization. These "Pictorial Structures" approaches assume the graph of the body is a tree which makes inference tractable [3, 13, 17].

The pictorial structures approach however has problems as illustrated in Figure 2 where multiple body parts explain the same image regions. The problems arise from the assumption that the global image likelihood can be expressed as a product of individual local terms (one per part), without regard to occlusion. To deal with this, previous algorithms have sampled multiple poses from the solution space and then used an external global likelihood to choose among the sampled alternatives [3]. Alternatively Ramanan and Forsyth [13] first find a solution for one side of the body and then remove the image regions explained by that solution from future consideration. They then solve for the other side of the body. While this sidesteps the problem it does not explicitly model the possible occlusion relationships and the algorithmic solution looses the probabilistic elegance present in the graphical model formulation.

Alternatively one can impose strong global constraints on the allowed poses that prohibit solutions like those in Figure 2 (*b*) and (*c*) [8]. This may be appropriate when the activity is known and the range of poses is highly constrained; for example, walking poses that can be represented using a small number of hidden variables [12]. We argue that these strong priors are invoked to deal with inadequate image likelihoods. In Figure 2 the local likelihoods prefer the *wrong* solutions and hence the prior is *fighting* with the likelihood to undo it's mistakes. Furthermore strong priors are unable to cope with unusual activities such as Figure 1.

The closest work to ours addresses the problem with the image likelihoods for 3D articulated hand pose estimation [21]. They explicitly modeled occlusions in 3D and dealt with distributed reasoning in graphical models using Nonparametric Belief Propagation [22]. The approach dealt with the issue of overcounting image evidence but did not address the problem of having the model explain as much of the image evidence as possible locally. They also dealt only with tracking from a hand initialized pose; here we go further to deal with automatic initialization. Our formulation allows for more general likelihoods, and outlines a competing inference algorithm that uses conditional distributions instead of potential functions as constraints between parts.

In summary we propose a method for approximating the global likelihood using local likelihoods. This allows us to use a part-based graphical model of the body and perform inference with a generic BP algorithm. Unlike [3] we deal with the continuous estimation of part locations, orientation, foreshortening and scale. Like previous approaches, for now we assume a known view but multiple views can be searched simultaneously and it is relatively straightforward to compare the results to select the best view. Without strong priors, the method finds solutions that better explain the image evidence.

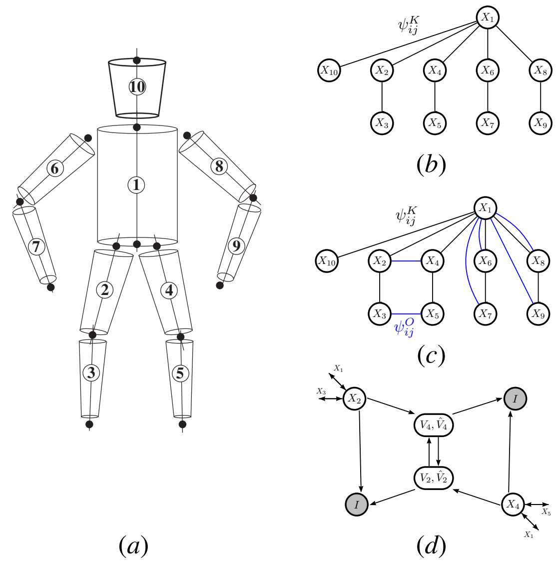


Figure 3. **Representing body as a graph**. Figure (*a*) shows the representation of the body as a graph with body parts labeled using the corresponding node numbers; (*b*) shows the corresponding tree-based representation of the body, and (*c*) our extended body model that contains additional occlusion constraints designated by edges in blue; (*d*) shows actual directed graphical model interactions encoded by a single blue edge in (*c*) between X_2 and X_4; I is the image evidence.

3. Modeling the body

The body is represented as a graphical model (Figure 3) in which nodes in the graph correspond to the rigid body parts and directed edges to the probabilistic constraints between parts encoded using conditional distributions. The pose of the body is $Y = \{X_1, X_2, ..., X_P\}$, where $X_i \in \mathbb{R}^5$ is a state of an individual articulated part i and P is the total number of such parts in the object. Each body part is modeled using a trapezoid in 2D, for which the state $X_i \in \mathbb{R}^5$ represents (x,y) position, rotation, scale and foreshortening in the image coordinate frame.

4. Likelihood

To estimate the pose of an object we must be able to evaluate how well different body configurations explain observed image data. We formalize this using a probabilistic likelihood function that takes a body pose and the image evidence and returns the likelihood of the pose. The desired properties of a good likelihood function lie in its robustness to partial occlusions, camera noise, changing lighting and the variability of appearance of the body.

4.1. Global vs. Local Image Likelihoods

Given the state of the body Y, we define a global likelihood $\phi(I|Y)$ in terms of some features I observed in an image. To support distributed modeling of the body we write this global likelihood as the product of local likelihood terms $\phi(I|Y) \propto \prod_{i \in \{1..P\}} \phi(I|X_i)$. Drawing inspiration from [3] and [23], we define local likelihoods in terms of the product of individual pixel likelihoods in sub-regions of the image that are defined by the local state X_i.

Formally, we assume that pixels in an image, I, can be partitioned into three disjoint sub-sets $\Omega_1(X_i) \cup \Omega_2(X_i) \cup \Omega_3(X_i) = \Upsilon = \{1..|I|\}$; where $\Omega_1(X_i)$ is the set of pixels enclosed by part i as define by the state X_i; $\Omega_2(X_i)$ contains the pixels outside part i that are statistically correlated with the part i (for example pixels in the border slightly outside the limb); and $\Omega_3(X_i) \equiv \Upsilon - (\Omega_1(X_i) \cup \Omega_2(X_i))$ which corresponds to the set of pixels where we assume no knowledge of the image statistics based on the pose of part i. Assuming pixel independence, we write the local likelihood $\phi(I|X_i)$ for the part i as a product of individual pixel probabilities as $\phi(I|X_i) =$

$$\prod_{u \in \Omega_1(X_i)} p_1(I_u) \prod_{s \in \Omega_2(X_i)} p_2(I_s) \prod_{r \in \Omega_3(X_i)} p_3(I_r) \qquad (1)$$

for pixels I_j, $j \in \Upsilon$.

The standard pictorial structures silhouette likelihood [3] can easily be written in this form, by letting I be a silhouette image obtained by background subtraction and by setting $p_1(I_u = 1) = q_1$, $p_2(I_s = 1) = q_2$, and $p_3(I_r = 1) = 0.5$ for some constants $0 \leq q_i \leq 1$ and $p_i(I_j = 0) = 1 - p_i(I_j = 1)$. For other non binary features such as limb/skin color we can express $p_1(I_u)$ and $p_2(I_s)$ as a ratio of learned foreground and background distributions; for example

$$p_{1,C}(I_u) = \frac{p_{skin}(I_u)}{p_{skin}(I_u) + p_{bkgd}(I_u)}$$

$$p_{2,C}(I_s) = \frac{p_{bkgd}(I_s)}{p_{skin}(I_s) + p_{bkgd}(I_s)}.$$

4.2. Occlusion-sensitive Local Likelihoods

The above formulation is only valid if the local terms $\phi(I|X_i)$ for $i \in \{1..P\}$ are independent. In absence of occlusions, this assumption holds and likelihoods factor. When limbs occlude each other however, the assumption does not hold and the product of local likelihoods gives a poor approximation to the global likelihood (see Figure 2).

To allow a similar decomposition (and hence distributed inference) when occlusions exist, we augment the state, X_i, of limb i with two sets of binary hidden variables $V_i = \{v_{i,u}\}$ and $\hat{V}_i = \{\hat{v}_{i,u}\}$, where u is a pixel $u \in \Upsilon$. Let $v_{i,u} = 0$ if pixel u for the part i is occluded by any other body part, and 1 otherwise. Intuitively this corresponds to the "visibility" of the part i at a given pixel u. Notice that if $v_{i,u} = 1$ for some pixel $u \in \Omega_1(X_i)$, then we know that part i at a given pose X_i generated pixel u in the image. Similarly we let $\hat{v}_{i,u} = 0$ if at pixel u for part i at X_i is occluding any other part, and 1 otherwise. Intuitively $\hat{V}_i$ encodes which pixels in the image could possibly be explained by other body parts that are further away from the camera. In particular if $v_{i,s} = 1$ and $\hat{v}_{i,s} = 1$ for a pixel slightly outside part i, $s \in \Omega_2(X_i)$, then that pixel, s, must have been generated by a background model (since by definition there cannot be any other part in front or behind i at s). Intuitively V_i and $\hat{V}_i$ in conjunction allow the likelihood to not only be sensitive to occlusions [21] but also to reason locally about globally plausible explanations of the image.

An illustration of these visibility variables is shown in Figure 4. For example, Fig. 4 (*c*) indicates that the torso is occluded by the lower arm ($v_{i,u} = 0$) and Fig. 4 (*g*) indicates that the arm is occluding part of the torso ($\hat{v}_{i,u} = 0$).

Modifying our likelihood, to take into account the hidden per-pixel binary occlusion variables we have

$$\phi(I|X_i, V_i, \hat{V}_i) = \prod_{u \in \Omega_1(X_i)} [p_1(I_u)]^{v_{i,u}} \prod_{s \in \Omega_2(X_i)} [p_2(I_s)]^{v_{i,s}\hat{v}_{i,s}} \prod_{r \in \Omega_3(X_i)} [p_3(I_r)]^{v_{i,r}\hat{v}_{i,r}} \quad (2)$$

Notice that $v_{i,u}$ and $\hat{v}_{i,u}$ are simply used as selectors. If pixel $u \in \Omega_1(X_i)$ is unoccluded then contribution of pixel u, $p_1(I_u)$, to the likelihood will be considered. Similarly, if pixel $s \in \Omega_2(X_1)$ is both unoccluded and unexplained then its contribution will be considered as well. Pixels for which $v_{i,u} = 0$ and/or $\hat{v}_{i,u} = 0$ will have constant likelihood 1.

The per-pixel occlusion-sensitive likelihoods are shown in Figure 4 for the torso (*e*) and lower arm (*h*). The local estimate of the global likelihood is simply the product of the pixel likelihoods where brighter indicates more likely.

It is important to note that conditioned on the sets of hidden variables V_i and $\hat{V}_i$ the local likelihoods $\phi(I|X_i, V_i, \hat{V}_i)$ are truly independent if V_i and $\hat{V}_i$ are consistent across all $i \in \{1..P\}$. By consistency here we mean that parts do not assume mutually occluding states for example (meaning that there may exist only one part i for which $v_{i,u} = 1$, for all others $v_{j,u} = 0$, where $j \in \{1..P\}/i$). This ensures that $\phi(I|Y) \propto \prod_{i \in \{1..P\}} \phi(I|X_i, V_i, \hat{V}_i)$ always holds.

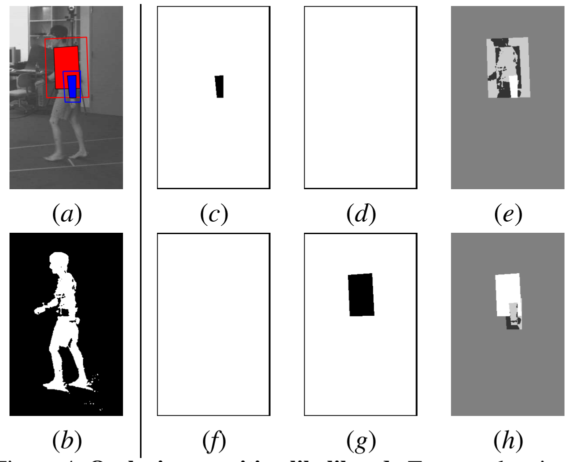


Figure 4. **Occlusion-sensitive likelihood**. Two overlapping parts (torso and lower arm) are shown in (*a*). The solid regions correspond to Ω_1 while the regions outside but enclosed by the line correspond to Ω_2. (*b*) shows the observed silhouette; (*c*) and (*f*) show the state of the hidden variables V_i for the torso and left lower arm respectively; (*d*) and (*g*) show the corresponding states of the $\hat{V}_i$'s; (*e*) and (*h*) shows the per pixel local occlusion-sensitive likelihoods with pixel brightness corresponding to high probability. Notice that in the cases where a part is both occluded and occluding other parts, both V_i and $\hat{V}_i$ will contain non-uniform structure.

5. Modeling Constraints

The body is represented by constraints between the parts that express traditional kinematic relationships as well as occlusion relationships between possibly occluding parts.

5.1. Occlusion Constraints

Enforcing the consistency of the hidden occlusion variables V_i and $\hat{V}_i$ requires reasoning that involves all potentially occluding and occluded parts for any given node i. We can express these occlusion constraints using pairwise conditional distributions $\psi^O_{ij}(X_j, V_j, \hat{V}_j | X_i, V_i, \hat{V}_i)$ between every pair of potentially occluding parts i and j. We formally encode the consistency of all occlusion relationships between part i and j using the unnormalized distribution:

$$\psi^O_{ij}(X_j, V_j, \hat{V}_j | X_i, V_i, \hat{V}_i) \propto \prod_{u \in \Upsilon} \begin{cases} 0 & \text{if } X_j \text{ occludes } X_i,\, u \in \Omega_1(X_j),\, v_{i,u} = 1 \\ 0 & \text{if } X_i \text{ occludes } X_j,\, u \in \Omega_1(X_i),\, v_{j,u} = 1 \\ 0 & \text{if } X_j \text{ occludes } X_i,\, u \in \Omega_1(X_i),\, \hat{v}_{j,u} = 1 \\ 0 & \text{if } X_i \text{ occludes } X_j,\, u \in \Omega_1(X_j),\, \hat{v}_{i,u} = 1 \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

Intuitively this simply enumerates all inconsistent cases and assigns them 0 probability. The first case for example can be interpreted as the following: if X_j occludes X_i and any pixel u is inside the image region of occluding part j, then $v_{i,u}$ corresponding to the visibility of the occluded part i at the pixel u must be set to 0.

5.2. Kinematic Constraints

Every pair of connected parts i, j in the body has an associated set of forward and backward kinematic constraints modeled as a Mixture of Gaussians, similar to [19] but in 2D. The kinematic conditional function, $\psi_{ij}^K(X_j|X_i)$, between parts i and j with corresponding states X_i and X_j is

$$\psi_{ij}^K(X_j|X_i) = \lambda_0 N(X_j; \mu_0, \Lambda_0) + (1-\lambda_0)\sum_{m=1}^{M_{ij}} \delta_{ijm} N(X_j; F_{ijm}(X_i), G_{ijm}(X_i)) \quad (4)$$

where λ_0 is a fixed outlier probability, μ_0 and Λ_0 are the mean and covariance of the Gaussian outlier process, and $F_{ijm}()$ and $G_{ijm}()$ are functions that return the mean and covariance of the m-th mixture component respectively; $\delta_{ijm} \geq 0$ are the weights of the mixture components and $\sum_{m=1}^{M_{ij}} \delta_{ijm} = 1$.

The conditional distributions were learned separately for 8-view based models using 3D motion capture data. The 3D body pose was projected into a desired camera view and the conditionals were learned from the 2D projections of individual limbs. We used a standard iterative Expectation-Maximization (EM) algorithm with K-means initialization for learning the Gaussian mixture model (GMM). All experiments in this paper used $M_{ij} = 8$ mixture components.

6. Inference

Inference in the standard pictorial structures model involves estimating the location and pose of every body part. With our occlusion-sensitive model we have the additional problem of dealing with the hidden occlusion variables. Given the formulation above, the joint probability for the graphical model with P body parts, can be written as

$$p(X_1, X_2, ..., X_P|I) \propto \sum_{V_i}\sum_{\hat{V}_i}[\prod_{ij}\psi_{ij}^K(X_j|X_i) \prod_{ij}\psi_{ij}^O(X_j, V_j, \hat{V}_j|X_i, V_i, \hat{V}_i)\prod_{j}\phi(I|X_j, V_j, \hat{V}_j)] \quad (5)$$

where X_i represents the state of the limb i; $\psi_{ij}^K(X_j|X_i)$ is the kinematic constraint compatibility term between the connected nodes i and j; $\psi_{ij}^O(X_j, V_j, \hat{V}_j|X_i, V_i, \hat{V}_i)$ is the occlusion compatibility between potentially occluding nodes i and j and $\phi(I|X_i, V_i, \hat{V}_i)$ is the local image likelihood. The two summations marginalize over the hidden occlusion variables in V_i and $\hat{V}_i$.

We solve for the part poses using belief propagation where the message update equations are:

$$m_{ij}^K(X_j) = \int_{X_i}\sum_{V_i}\sum_{\hat{V}_i}[\psi_{ij}^K(X_j|X_i) \phi(I|X_i, V_i, \hat{V}_i)\prod_{k\in A/j} m_{ki}^K(X_i) m_{ki}^O(X_i, V_i, \hat{V}_i)] \quad (6)$$

$$m_{ij}^O(X_j, V_j, \hat{V}_j) = \int_{X_i}\sum_{V_i}\sum_{\hat{V}_i}[\psi_{ij}^O(X_j, V_j, \hat{V}_j|X_i, V_i, \hat{V}_i) \phi(I|X_i, V_i, \hat{V}_i)\prod_{k\in A/j} m_{ki}^K(X_i) m_{ki}^O(X_i, V_i, \hat{V}_i)] \quad (7)$$

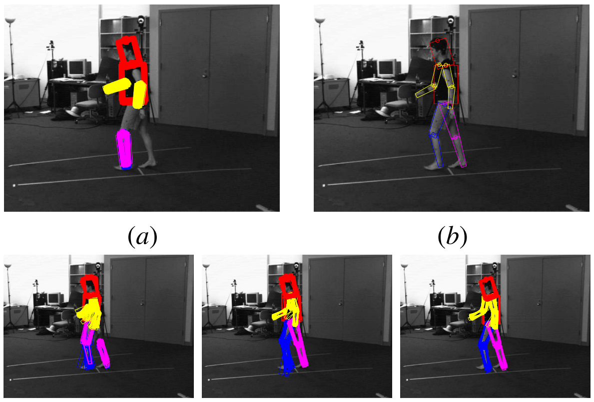
Inferring the state of the 2D body in our graphical model representation corresponds to estimating the belief (marginal) at each node in a graph, $b_i(X_i) =$

$$\sum_{V_i}\sum_{\hat{V}_i}\phi(I|X_i, V_i, \hat{V}_i)\prod_{k\in A} m_{ki}^K(X_i) m_{ki}^O(X_i, V_i, \hat{V}_i).$$

We use a form of continuous non-parametric belief propagation (PAMPAS) [6] to deal with this task. The messages are approximated using a kernel density formed by propagating particles through a conditional density [19]. In all the experiments we used 100 particles which, when propagated through the conditionals represented by mixtures of 8 Gaussians, resulted in density representations for the messages with 800 kernels; from products of these messages we sampled 100 particles. We modify the method to include an annealing step [2] with each iteration of PAMPAS that gradually introduces the effects of peaks in our local likelihoods; this modification is not essential. For the details on how the message updates are carried out using stratified sampling from the products of messages and a static proposal distribution see [19]. The illustration of the inference using PAMPAS with occlusion-sensitive likelihoods can be seen in Figure 5.

6.1. Message Updating for Occlusion Messages

It is intractable to sample occlusion variables V_i and $\hat{V}_i$ due to the exponentially large number of possible occlusion mask configurations. Consequently we approximate the computation of marginals using an analytic procedure introduced in [21]. Assuming we know depth ordering for the parts in a given view we compute the approximate message $m_{ij}^O(X_j, V_j, \hat{V}_j)$ for V_j and $\hat{V}_j$ explicitly. To do so, we must consider two cases: (1) where X_j is occluded by X_i and (2) where X_j is occluding X_i. We assume that potentially occluding parts have a known and unchanging depth order to simplify the formulation. In general, we could introduce an additional discrete hidden variable designating the depth order between parts and marginalize over it as well which would lead to a more complex inference scheme.



(a) (b)

(c)

Figure 5. **Occlusion-sensitive inference**. Figure (*a*) shows the proposal distributions for the six body parts drawn from ground truth pose and corrupted by Gaussian noise. Both left and right calves are initialized intentionally incorrectly on the left calf in the image; (*b*) shows the mean of the marginal distribution for each part after 3 iterations of belief propagation (BP). Figure (*c*) shows 100 samples from the marginal distributions after one, two and three iterations of BP. Notice that we initialize from a local maximum of the traditional likelihood function, precisely the place where most algorithms get "stuck", yet our algorithm is still able to recover the correct pose.

If X_j is occluded by X_i the message from X_i to X_j about the state of $\hat{V}_j$ is uninformative and can be written in terms of individual per-pixel hidden binary variables as $m^O_{ij}(\hat{v}_{j,u} = 1) = 1$, where $u \in \Upsilon$. The message for V_j is informative however and can be approximately computed as $m^O_{ij}(v_{j,u} = 1) \propto 1 - p(u \in \Omega_1(X_i))$, where $p(u \in \Omega_1(X_i))$ is simply the probability of pixel $u \in \Upsilon$ being inside the projection of X_i. Similar expressions can be derived for the case where X_j is occluding X_i, but are omitted for conciseness.

We can now approximate the marginal probability of a pixel u being "visible" for part j, $p(v_{j,u} = 1)$, by taking a product over all potential occluders,

$$p(v_{j,u} = 1) \propto \prod_i m^O_{ij}(v_{j,u} = 1). \tag{8}$$

Since $v_{j,u}$ is binary, the occlusion probability is simply $p(v_{j,u} = 0) = 1 - p(v_{j,u} = 1)$. Similarly for $p(\hat{v}_{j,u} = 1) \propto \prod_i m^O_{ij}(\hat{v}_{j,u} = 1)$, where $p(\hat{v}_{j,u} = 1)$ is the marginal probability of the pixel u not being explained by any other part i that is behind part j (further away from the camera). Computation of these marginals amount to "projecting" the distribution (represented in terms of weighted particles) for every possible occluder X_i into the image and summing over the resulting binary masks (with normalization).

We can now re-write the likelihood functions in terms of the marginal probabilities $z_{j,u} \equiv p(v_{j,u} = 1)$ and $\hat{z}_{j,u} \equiv p(\hat{v}_{j,u} = 1)$,

$$\begin{aligned} p(I|X_j, V_j, \hat{V}_j) = & \\ & \prod_{u \in \Omega_1(X_j)} [(1 - z_{j,u}) + z_{j,u} p_1(I_u)] \\ & \prod_{s \in \Omega_2(X_j)} [(1 - z_{j,s}\hat{z}_{j,s}) + z_{j,s}\hat{z}_{j,s} p_2(I_s)] \\ & \prod_{r \in \Omega_3(X_j)} [(1 - z_{j,r}\hat{z}_{j,r}) + z_{j,r}\hat{z}_{j,r} p_3(I_r)] \,. \end{aligned} \tag{9}$$

This equation downweights the image evidence for the part j at a pixel $u \in \Omega_1(X_j)$ as the probability of that pixel's visibility decreases (occlusion probability increases). Similarly, it also downweights the image evidence at the pixel $s \in \Omega_2(X_j)$ as the probability of that pixel being explained by another body part further away from the camera increases. Notice that this likelihood can be implemented efficiently by only considering regions of the image $\Omega_1(X_j)$ and $\Omega_2(X_j)$ for a given X_j, and precomputing $\prod_{r \in \Upsilon} [(1 - z_{j,r}\hat{z}_{j,r}) + z_{j,r}\hat{z}_{j,r} p_3(I_r)]$.

6.2. Limb Proposals

Plausible poses/states for some or all the body parts are needed as proposals [19]. There exist a number of efficient methods for detecting body parts in an image [9, 11, 16]. Here we took a simple approach and constructed a set of proposals by coarsely discretizing the state space and evaluating local part-based likelihood functions at these discrete locations. For all of the experiments here we discretized the state space into 5 scales, 5 foreshortenings, 20 vertical and 20 horizontal positions and 8 rotations. We chose the 100 most likely states for each part and used these as a particle based proposal distribution for belief propagation. It is important to note that not all parts need to be detected and, in fact, detecting all the parts is largely impossible due to the self occlusions. An example of the synthetic proposals for various parts of the body are shown in Fig. 5 (*a*). To initialize the search we used proposals for 6 parts: torso, head and four outermost extremities. All other parts were initialized with a uniform distribution over the entire state space.

7. Experiments

We learned occlusion-sensitive models for 8 discrete views of a person including frontal, side and 3/4 views. For each view we assume the depth ordering of the body parts is known. In all experiments the likelihood uses a combination of silhouette and color/intensity information (assuming independence). For the silhouette likelihood we used the pictorial structures model and learned $p_{1,S}(I_u = 1) = q_1$ and $p_{2,S}(I_s = 1) = q_2$ using the procedure described in [3]. Similar to [3] we assumed that $p_{3,S}(I_r = 1) = 0.5$. For the

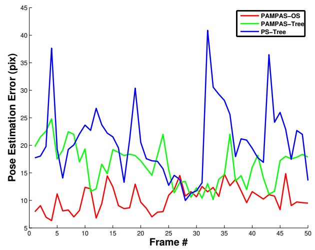


Figure 6. **Quantitative Performance Evaluation**. Mean error of the joint locations for each frame of 50 frame image sequence with ground truth [19]. For the description of the metric see text.

	Strong Prior	Discrete State	Mean Error	Std
PAMPAS-OS	No	No	10.33	2.25
PAMPAS-Tree	No	No	16.40	3.67
PS-Tree	No	Yes	20.84	6.64
PS-Tree [8]	No	Yes	13.79	3.99
LBP [8]	Yes	Yes	12.00	3.99
Factor [8]	Yes	Yes	6.42	1.55

Figure 7. **Overall Performance Comparison**. Performance of the occlusion-sensitive inference compared with two tree-based algorithms implemented by us. We also compare to the results reported by [8] on the same image sequence.

color/intensity likelihood we learned a kernel density model for each part and the background.

For frontal views, the lack of self occlusion means that tree based approaches will usually perform well. Consequently we focus on the more challenging side-views containing occlusion. We quantitatively compare our approach (PAMPAS-OS) to leading tree-based methods using 50 frames from the Brown ground truth sequence [19] using the metric presented by [8] which computes the average distance error between a set of 15 estimated marker locations corresponding to the joints.

For comparison we implemented two tree-based methods: pictorial structures (PS-Tree) [3] and a variant of our approach that does not model occlusions (PAMPAS-Tree) by simply removing the occlusion constraints from our model. Figure 7 shows the mean error for 15 markers at every frame for the three methods. Following [8] we deal with the left/right ambiguity by switching the left/right limbs and reporting the interpretation with a smallest error.

Our occlusion-sensitive inference approach outperforms pictorial structures by 50% (25% for the implementation in [8][1]). We found that occlusion-reasoning accounts for a 37% performance gain over the simple PAMPAS-Tree method. According to the published literature [8] our approach also outperforms max-product loopy-BP, but does not do as well as the common-factor model (Factor) presented in [8]. This is not surprising, since the common-factor model uses a walking prior learned for this data. Our approach does not assume a strong prior on the motion, and hence is not restricted to any given motion type.

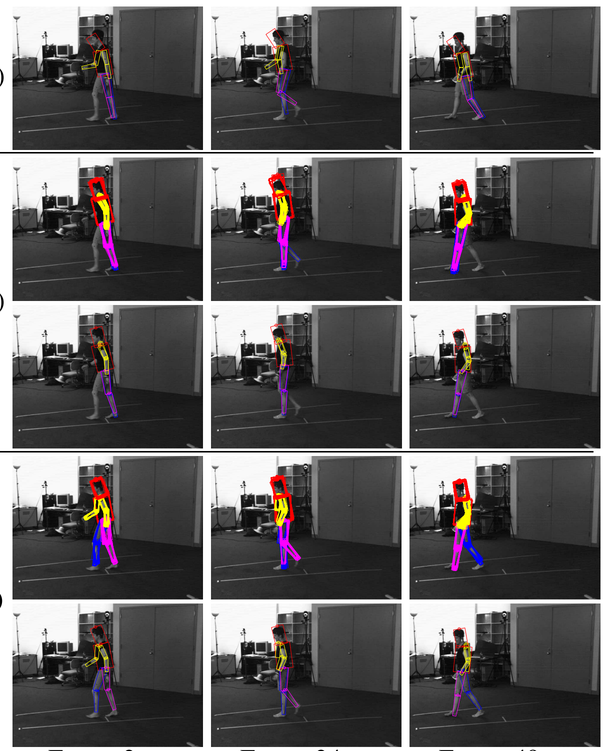
Figure 8 illustrates the behavior of PS-Tree, PAMPAS-Tree and PAMPAS-OS on a few frames of the sequence. As expected we observed many failures in the pictorial structures model due to the overlapping parts. PAMPAS-Tree, not surprisingly had similar modes of failure while the occlusion-sensitive PAMPAS-OS does a better job of explaining the image evidence.

In addition to the quantitative sequence we also ran our model on less structured scenarios from TV and movies for

(*a*)

(*b*)

(*c*)



Frame 2 Frame 24 Frame 49

Figure 8. **Visual Performance Evaluation**. (*a*) MAP estimates for the tree-based implementation of pictorial structures on three frames from our test sequence. Performance of occlusion-insensitive and occlusion-sensitive PAMPAS is shown in (*b*) and (*c*) respectively. The top rows show 100 samples from the marginal distribution at every node (belief) after 5 iterations of BP, and bottom rows the weighted mean computed over those samples.

which strong prior models will not work. Figure 9 illustrates two representative results. In both cases, camera motion makes background subtraction difficult. Crude background subtraction was obtained using homographies estimated between 2 frames sufficiently far apart in time (using the code from `http://www.robots.ox.ac.uk/~vgg/`). Color likelihoods were defined as in [13].

Our current un-optimized implementation of PAMPAS-

[1] Our independent implementation of PS-Tree [3] resulted in somewhat larger error than reported in [8].

(*a*)

(*b*)

Figure 9. **Occlusion-sensitive reasoning in movies**. Results on frames from TV/films. Left column shows 100 samples from the marginal distribution (belief) after 3 iterations of BP, and right column shows the weighed mean pose.

OS in Matlab takes roughly 5 minutes for message passing, and 1.5 minutes for belief estimation per frame. The occlusion constraints account for a 43% overhead over PAMPAS-Tree.

8. Summary and Conclusions

We introduce a novel approach for articulated 2D body pose estimation that uses occlusion-sensitive local image likelihoods that approximate the global likelihood by accounting for occlusions and competing explanations of image evidence by multiple parts. We model occlusion relationships between parts explicitly by introducing two sets of per-pixel hidden binary variables for each part. The resulting occlusion reasoning involves interactions between non-adjacent parts which introduces loops in the graphical model representation of the body. To achieve tractable real-valued inference in such a graph, we also introduced an extension to the approximate belief propagation inference algorithm (PAMPAS) that takes into account, and analytically marginalizes over, the hidden occlusion variables of our model.

We quantitatively compare our approach to two state-of-the-art algorithms using tree-structured kinematic models, as well as to published results in the literature. The proposed approach performs favorably and solves the problem of competing models that tend to match multiple body parts to the same image evidence without the addition of strong priors. Experimental results illustrate that our model has pose error at least 25% lower than tree-structured models. We also show that our approach performs favorably in complex scenarios, where strong assumptions about the kinematic motion of the body are not appropriate.

Acknowledgments. This work was partially supported by Intel Corporation. Leonid Sigal was also supported by NSF IGERT award #9870676. We would like to thank Andrew Zisserman for providing the data from movie 'Run Lola Run'.

References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, pp. 8–15, 1998.

[2] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2004.

[3] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, (61)1:55-79, Jan. 2005.

[4] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstruction of (3D) human motion from single-camera video. *NIPS*, pp. 820–826, 1999.

[5] G. Hua, M.-H. Yang, and Y. Wu. Learning to estimate human poses with data driven belief propagation. *CVPR*, vol. 2, pp. 747–754, 2005.

[6] M. Isard. Pampas: Real-valued graphical models for computer vision. *CVPR*, vol. 1, pp. 613–620, 2003.

[7] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parametrized model of articulated motion. *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[8] X. Lan and D. Huttenlocher. Beyond trees: Common factor models for 2D human pose recovery. *ICCV*, vol. 1, pp. 470–477, 2005.

[9] M. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. *CVPR*, vol. 2, pp. 334–341, 2004.

[10] M. Lee and I. Cohen. Human upper body pose estimation in static images. *ECCV*, vol. 2, pp. 126–138, 2004.

[11] G. Mori, X. Ren, A. Efros and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *CVPR* vol. 2, pp. 326–333, 2004.

[12] D. Ormoneit, M. Black, T. Hastie, and H. Kjellström. Representing cyclic human motion using functional analysis. *Image and Vision Computing*, 23(14):1264–1276, Dec. 2005.

[13] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *CVPR*, vol. 1, pp. 271–278, 2005.

[14] D. Ramanan and D. Forsyth, Finding and tracking people from the bottom up. *CVPR*, vol. 2, pp. 467–474, 2003

[15] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. *ICCV*, pp. 824–831, 2005.

[16] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. *ECCV*, vol. 4, pp. 291–303, 2004.

[17] R. Ronfard, C. Schmid and B. Triggs. Learning to parse pictures of people . *ECCV*, vol. 4, pp. 700–714, 2002.

[18] H. Sidenbladh, M. Black and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, vol. 2, pp. 702–718, 2000.

[19] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. *CVPR*, vol. 1, pp. 421–428, 2004.

[20] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *IJRR*, 22(6), pp. 371–391, 2003.

[21] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. *NIPS*, pp. 1369–1376, 2004.

[22] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. *CVPR*, vol. 1, pp. 605–612, 2003.

[23] T. Zhao and R. Nevatia. Bayesian Human Segmentation in Crowded Situations. *CVPR*, vol. 2, pp. 459–466, 2003.