Constraint Integration for Efficient Multiview Pose Estimation with Self-Occlusions

Abhinav Gupta
University of Maryland, College Park
agupta@cs.umd.edu

Abstract

Detection and tracking of articulated objects such as humans is an important task in computer vision. While significant advancement has been made on this problem recently, many limitations remain due to the complexity in handling many of the constraints present in the problem. In this paper, we present a system that incorporates a variety of "new" constraints in a unified multi-view framework to automatically detect and track articulated objects in possibly crowded scenes. These constraints include the occlusion of one part by another and the high correlation between the appearance of certain parts (the two arms, for instance). The graphical structure (non-tree) obtained is optimized in a nonparametric belief propagation framework. Efficient methods are presented in order to reduce the complexity of the problem.

1. Introduction

Detection and tracking of articulated objects such as humans in crowded scenes is an important, albeit unsolved problem in computer vision. The problem is hard because of occlusions, a high dimensional problem space and high variability in the appearance of humans due to body shape and clothing. Most prior work has focussed solely on tracking, where the initialization is given[3, 17]. Recently, there has been a focus on automatic detection of body pose that could then be used to initialize/re-initialize tracking systems[5, 10, 18].

There are a wide range of approaches to human pose estimation. Much of the work model the human body as a tree structure. Here, each part is represented by a node in the tree and there is an edge joining the parts between which there are kinematic relations. The edges impose constraints on the possible locations of different parts. These constraints may be applied either in 2D [4, 13] or 3D [19]. Felzenszwalb et. al. [5] presented a deterministic linear time algorithm using dynamic programming to solve for the best pose configuration in such tree structures. Other optimization approaches like Data Driven Belief Propagation [6] and Markov Chain Monte Carlo algorithm [10] have also been used to estimate the probability distributions of the locations of body parts.

However, there are limitations to a tree structure. Kinematic relations between parts that are not connected to each other cannot be represented. Furthermore, occlusion of one part by another cannot be modeled nor can the constraint due to the high correlation between the appearance of certain parts (e.g. the two hands)[12]. There has been some recent work to overcome these limitations. Lan et. al [9] use factor graphs to add constraints like the balance of a body while walking; Ren et. al [14] use Integer Quadratic Programming (IQP) to add pairwise constraints such as similarity in the appearance of left and right body-parts.

Ioffe et. al [7] proposed using a mixture of trees to handle occlusions. The mixture includes all possible trees resulting from removing nodes from the base tree under different occlusion scenarios. However, modeling the conditionals between non-connected parts is very difficult; it does not provide very strong constraints, leading to false part localizations. At the same time, the problem space becomes very large due to the need to optimize over the entire ensemble of trees.

The problem can be simplified by assuming that one can segment the person, say using background subtraction [2, 9, 11]. While this reduces the search space significantly, these approaches generally do not handle self-occlusion or people occluding one another.

A complementary approach [1, 16] is to learn pose configurations from training images and sequences. Like all appearance based techniques, they have difficulty generalizing to new views or unconventional poses.

In this paper, a multiple camera based approach for estimating the 3D pose of humans in a crowded scene is presented. The system incorporates a variety of constraints, including the occlusion of one part by another and appearance consistency across parts, in a unified framework. Inclusion of these constraints, however, breaks the tree structure of the graphical model. Consequently, the optimization becomes quite complex, and particle-based belief propagation is utilized to optimize over the space of possible body configurations and appearances. Messages encoding different

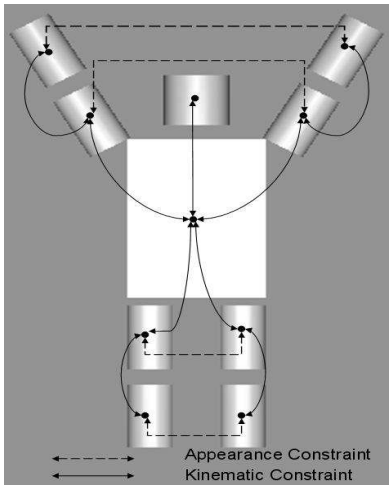


Figure 1. The object model used for human beings. The solid lines represent edges in set E_1 and dashed lines represent edges in set E_2. We have not shown occlusion edges in the above graph. Every part is connected to all other parts by occlusion constraint edge.

constraints are passed between parts not directly connected in a tree structure. Several constructs are also introduced to efficiently prune the search and locate parts.

The paper is organized as follows. In section 2 we discuss our human body model followed by a discussion on how to pass information between the parts in section 3. Section 4 provides a description of visibility analysis and likelihood computations. We provide a system overview and extend it to tracking poses in sections 5 and 6 respectively. Finally, results are presented in section 7 before concluding in section 8.

2. Modeling the Human Body and Problem Formulation

Our 3D human body model (Figure 1) consists of $n = 10$ body parts (head, torso, left upper arm etc.). Each body part (except the torso which is modeled as a cuboid) is modeled as a cylinder and is represented by a node in a graph. This represents a random vector $\Phi_i = (l_i, \psi_i)$, where l_i and ψ_i represent the location and appearance parameters of part i respectively. The location of each part, l_i, is further parameterized by $l_i = (l_i^s, l_i^e)$ where l_i^s is the 3D position of the starting point of the limb and l_i^e is the 3D position of the ending point of the limb.

The nodes of the graph are connected by three types of edges. The first enforces kinematic constraints between parts. To obtain a tree model, like those typically used in the literature, one would only connect parts using edges of this kind. The second type of edge represents appearance constraints which are introduced by the symmetry of left and right body part appearances. The third type of edge represents occlusion constraints across parts that can occlude each other. The model is represented by $\theta = (E_1, E_2, E_3, c^1, c^2, c^3)$, where the set of edges E_1, E_2 and E_3 indicates which parts are connected by edges of the first, second and third type respectively; c^1, c^2 and c^3 are the connection parameters for these edges.

We need to find the probability distribution of the pose configuration of a human body, given by $L \equiv (\Phi_1, \Phi_2 ..... \Phi_n)$. In an M camera setup, if I_j denotes the image from the j^{th} camera, then $P(I_1 ....I_M|L)$ is the likelihood of observing the set of images given the 3D locations and appearances of the body parts. The distribution of $P(L)$ is the prior over the possible configurations. The goal is to maximize the posterior distribution, $P(L|I_1 ....I_M)$, which measures the probability of a particular configuration of the human body given M views and the object model. Using Bayes' rule,

$$P(L|I_1 ....I_M) \propto P(I_1 ....I_M|L)P(L) \tag{1}$$

Assuming that the location and appearance priors are independent of each other, the prior distribution $P(L)$ is

$$P(L) = P(l_1 .....l_n)P(\psi_1 .....\psi_n) \tag{2}$$

The prior distribution over the object part locations and appearances are modeled by two separate Markov random fields with edge sets E_1 and E_2. The joint distribution for the tree-structured prior defined by E_1 can be expressed as:

$$P(l_1, l_2 ...l_n) = \frac{\prod_{(v_i,v_j)\in E_1} P(l_i, l_j)}{\prod_{v_i \in V} p(l_i)^{deg(v_i)-1}} \tag{3}$$

where V is the set of nodes in the graph and $deg(v_i)$ is the degree of vertex, v_i, in the tree, $G = (V, E_1)$ (subgraph consisting of edges in E_1 only). A similar expression can be written for $P(\psi_1, \psi_2 ....., \psi_n)$. Since any absolute location or appearance is not preferred over another, the terms representing the priors for single part locations can be neglected. Furthermore, as in most prior work[9, 18], potential functions rather than distributions are used to avoid normalization computations. Then, one obtains:

$$P(l_1, l_2 .....l_n) \propto \prod_{(v_i,v_j)\in E_1} \varphi_{ij}(l_i, l_j) \tag{4}$$

$$P(\psi_1, \psi_2 ....., \psi_n) \propto \prod_{(v_i,v_j)\in E_2} \phi_{ij}(\psi_i, \psi_j) \tag{5}$$

where φ_{ij} and ϕ_{ij} are the potential functions over the clique.

For articulated objects, pair of parts are connected by flexible joints. Ideally, the distance between the ending-point of the first part and the starting point of the second connected part in 3D should be zero. Thus, the clique potential for a pair of parts, connected by edges in E_1, can be modeled as:

$$\varphi_{ij}(l_i, l_j) = N(d(l_i, l_j), 0, \sigma_{ij}^1) \quad (6)$$

where $d(l_i, l_j)$ denotes the euclidean distance between the points l_i^e and l_j^s.

For appearance constraints, let $D(\psi_i, \psi_j)$ denote the distance between two appearance vectors. Ideally, the distance should be zero, assuming left and right body parts have similar appearance. The appearance potential, ϕ_{ij}, is modeled as:

$$\phi_{ij}(\psi_i, \psi_j) = N(D(\psi_i, \psi_j), 0, \sigma_{ij}^2) \quad (7)$$

Section 4.2 discusses how part appearances are modeled and how the distance ,$D(\psi_i, \psi_j)$, is computed.

The computation of the likelihood $P(I_1....I_M|L)$ is tricky due to the consideration of occlusion. The imaging of every camera is modeled as conditionally independent processes. Similarly, the observation of different parts is assumed to be conditionally independent. This allows us to decompose the likelihood as:

$$P(I_1....I_M|L) \propto \prod_{i=1}^{n}\prod_{j=1}^{M} P_i(I_j|l_1...l_n, \psi_i) \quad (8)$$

Note that, due to the possibility of occlusion, the likelihood of each part depends not only on the position of the part, but also on the positions of other parts. While one may be able to use the likelihood in this form in tracking applications, using it for automatic "detection" is prohibitively expensive. To overcome this, we could introduce a new set of binary 'visibility' variables $v_i^j(l_i)$, that refer to the visibility of a part i at location l_i from camera j. While these visibility variables would be dependent upon the position of all other parts, the likelihood for part i would be independent of the location of other parts if its visibility were given. Then, one could write the likelihood, $P(I_1....I_M|L)$, as:

$$\prod_{i=1}^{n}\prod_{j=1}^{M}\sum_{v_i^j\in\{T,F\}} P_i(I_j|l_i, v_i^j(l_i))P(v_i^j(l_i)|l_1....l_{i-1}, l_{i+1}....l_n) \quad (9)$$

The term $P_i(I_j|l_i, v_i^j(l_i) = TRUE)$ represents the likelihood of observing the image from camera j given that the part is visible from this camera while $P_i(I_j|l_i, v_i^j(l_i) = FALSE)$ represents the likelihood of observing the image given that the part is occluded from the camera. However, parts may be partially visible in which case $v_i^j(l_i)$ is neither true nor false. To approximate this, $v_i^j(l_i)$ is defined as the visibility of a random point on the skeleton of the part. In Section 4.1, we will discuss how to compute the visibility variables and in section 4.3, we will discuss in more detail how to compute the likelihoods.

3. Particle Based Belief Propagation

In the previous section, a graphical model for human body parts was developed. In order to solve for the best configuration in such a graphical model, the framework proposed in [18] can be utilized. Essentially, the system optimizes for the posterior of each part and the interactions between different parts are handled via messages in a non-parametric belief propagation framework. A variant of the PAMPAS algorithm is used for non-parametric belief propagation [8]. The framework provides a natural approach for enforcing constraints across parts, including those of occlusion and appearance matching.

There are, essentially, three kinds of unknowns that need to be estimated simultaneously: the location of each part, the appearance of each part and the visibility variables. The probability densities of part location and appearance are represented via monte carlo particles while visibility variables can be computed from probabilistic occlusion maps.

The following messages are used to pass information to a part:

- The locations of neighboring connected body parts (eg. location of lower left leg and torso particles is passed to upper left leg). These location are used to apply kinematic constraints.
- The appearance of the corresponding symmetric part (eg. appearance of right upper leg is passed to the left upper leg).
- The visibility information from other parts that may occlude this part (eg. upper left leg receives the occlusion map from all other parts in order to update its likelihood distribution)

At iteration r, a message m_{ij} from node i to j along an edge in E_1 or E_2 may be represented as:

$$m_{ij}^r(\Phi_j) = \int \varphi_{ij}(l_i, l_j)\phi_{ij}(\psi_i, \psi_j)P_i(I_1....I_M|L) \prod_{k\in E_1\setminus j} m_{ki}^{r-1}(\Phi_i) \prod_{o\in E_2\setminus j} m_{oi}^{r-1}(\Phi_i)dl_i$$

where $\mathbf{v}_i = (v_i^1, ..., v_i^M)$. Note that $\varphi_{ij}(l_i, l_j) = 1$ for messages along edges in E_2 and $\phi_{ij}(\psi_i, \psi_j) = 1$ for messages along edges in E_1. Messages along E_3 alter the visibility variables:

$$m_{ij}^r(\mathbf{v}_j) = \int Occl(l_i)P_i(I_1....I_M|L) \prod_{k\in E_1\setminus j} m_{ki}^{r-1}(\Phi_i) \prod_{o\in E_2\setminus j} m_{oi}^{r-1}(\Phi_i)dl_i$$

where $Occl(l_i)$ defines the occluding characteristics of part l_i and affects the visibility parameters of part j.

Then, the posterior distribution of a body-part $Pos^r(\Phi_i)$ can be computed as:

$$Pos^r(\Phi_i) \propto P_i(I_1....I_M|L) \prod_{k\in E_1\setminus j} m^r_{ki}(\Phi_i) \prod_{o\in E_2\setminus j} m^r_{oi}(\Phi_i) \quad (10)$$

To initialize the system, uniform appearance priors and full visibility of each part is used; that is, it is assumed that all parts are fully visible. At any iteration, the posterior distribution of each part is approximated by a set of particles which are sampled using importance sampling. The set of these particles is used to generate the messages to be passed along appropriate edges in order to enforce inter-part relationships. Updating the parameters for the different parts in turn, the method eventually leads to a stable parameter estimation after several iterations. The particle-based belief propagation is especially effective since the probability distributions are typically not gaussian in nature, especially in the initial iterations, and hence using any parametric model would lead to a loss of information.

4. Computing Priors and Likelihoods

4.1. Computing Part Visibility

We discuss how to compute $P(v^j_i(l_i)|l_1..l_{i-1}, l_{i+1}, ..l_n)$, which represents the probability of visibility of a random point of the skeleton of part i in view j, given the pdf's of locations of parts $l_1 \ldots l_n$. If the exact positions of parts in 3D were known, computing $P(v^j_i(l_i)|l_1..l_{i-1}, l_{i+1}, ..l_n)$ would be straightforward. However, only the posterior distributions of the locations of the parts after the previous iteration are known. To compute the probability, notice that a part is not occluded if and only if it is not occluded by any of the parts, allowing us to utilize an independence relation between the occlusion from different parts. Thus, the probability of visibility of a part i in view j, $P(v^j_i(l_i)|l_1..l_{i-1}, l_{i+1}..l_n)$ represented by Pv^j_i, can be broken down into product of probability of visibilities from different parts as:

$$\begin{aligned} Pv^j_i &= \prod_{k=1,2..i-1,i+1...n} P(v^j_{ik}(l_i)|l_1..l_{i-1}, l_{i+1}..l_n) \\ &= \prod_{k=1,2..i-1,i+1...n} P(v^j_{ik}(l_i)|l_k) \end{aligned} \quad (11)$$

The above equation requires us to compute $P(v^j_{ik}(l_i)|l_k)$, which represents the probability that a part i is not occluded by a part k.

To compute this probability efficiently, "occlusion maps" are introduced. An occlusion map of a part k, $O^j_k(x, y, z)$, stores the probability that a 3D point (x, y, z) will be occluded by part k in view j (Figure 2 illustrates an occlusion map of a sphere). The occlusion map of a body part depends on the shape and location of the part. The occlusion maps have to be updated at every iteration because the probability distribution of location of each part changes after each iteration. For updating the occlusion map of part k, the region of occlusion[1] for each particle of k is computed. The update is made using the following equation:

$$O^{r+1,j}_k(x, y, z) = \frac{n_{occ}}{n} \quad (12)$$

where r is the iteration number, n_{occ} is the number of particles that support the fact that a point (x, y, z) will be occluded by part k in view j, and n is the total number of particles used for computing the message. Intuitively, the probability that a 3D point (x, y, z) is occluded by part k is proportional to the number of particles of part k that occlude the point.

To provide smoother updates to the occlusion maps and handle errors in approximating the probability calculations, it is useful to update the occlusion maps incrementally:

$$O^{r+1,j}_k(x, y, z) = (1 - \beta)O^{r,j}_k(x, y, z) + \beta(\frac{n_{occ}}{n}) \quad (13)$$

where β determines the rate of change of the occlusion maps ($\beta = 0.2$ was used in our experiments).

Using the occlusion map of part k for view j, the probability of visibility of a point object i at location, $l_i = (x, y, z)$ in view j, can be computed as:

$$P(v^j_{ik}(l_i)|l_k) = 1 - O^j_k(x, y, z) \quad (14)$$

In order to address the finite size of the part, $P(v^j_i(l_i)|l_k)$ is approximated by averaging the different visibility probabilities along the part skeleton.

4.2. Part Appearance

The appearance of a part is modeled by computing its color as a function of height. A single color model fails to capture the color variation along the part axis. A histogram would be too expensive to compute for all the hypothesis and is thus not used. Thus, the appearance of a part can be represented by a vector that contains n_1 different color vectors along the part. The euclidean distance is used to compute the distance between two appearance vectors.

4.3. Image Likelihoods

Each body part is modeled as a cylinder. Under orthographic projection, the image of a cylinder will consist of parallel lines for two occluding contours of the part, except the two circular surfaces at the joints which are normally not

[1]The region of occlusion is the 3D region that will be occluded by the part

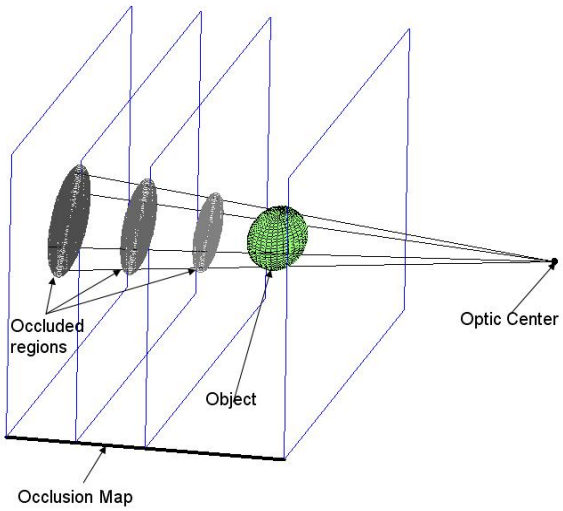


Figure 2. The occlusion map created by a sphere. The cone behind the sphere is the region of occlusion in 3D. The probability of visibility is decreased for every 3D point lying within the cone.

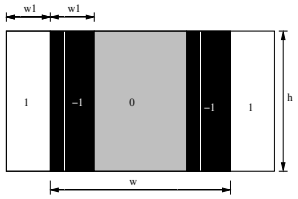


Figure 3. The filter used for finding image likelihoods for parallel lines. w represents the projected width of the body part and h represents the height of the part. The grey portion represents the part of image that will not be considered in computing the response. The white, black and grey portions have weights 1,-1 and 0 respectively.

detectable. The response of a filter shown in Figure 3 is used to find such parallel lines. The filter gives high response for parallel lines separated by distance w and is robust to moderate deviation from the parallel line assumption.

An exponential dependence of the likelihood on the filter response is employed, so the likelihood of the image given that the object-part is visible from the camera is:

$$P_i(I_j|l_i, v_i^j(l_i) = TRUE) \propto e^{(1-resp(l_i^j))} \qquad (15)$$

where l_i^j is the location where part i will be projected in image j. More complicated models and filters can also be used[15]. Computation of $P_i(I_j|l_i, v_i^j(l_i) = FALSE)$ represents the case when the part is occluded. It can also be treated as computing the likelihood of observing a random pattern at location l_i^j with no preference given to one pattern over another [2]. Therefore, the likelihood can be assigned a fixed constant in this case.

5. Efficiency Considerations and System Overview

We discuss some additional features used to make the system fast and fully automatic.

Our method requires the computation of the posterior distribution for each part in the graphical model. The computation of such a distribution, however, can be prohibitively expensive since it requires search over a large configuration space. In order to perform this search efficiently, two methods were studied. In the first, the space is first pruned via priors, while in the second method, it is pruned via likelihoods. In the first method, the high likelihood parameters of previous parts and anthropometric data are used to prune the search region for a part in 3D. For example, after finding the probability distribution of the upper arm, one can prune the search area in 3D for the lower arm.

However, the process is too expensive since there are many cases in which the search space cannot be sufficiently constrained (the search space becomes especially large for the four end limbs, for instance). For such cases, likelihood-based search in 2D is used, which finds the possible limbs in each 2D image using the responses of the filter. First, a search region in 2D is determined based on the positions of the previous parts. Then, regions that give a high filter response are identified in each image. These high likelihood part hypotheses in 2D are then matched across views using epipolar constraints. Searching along epipolar lines for the starting and ending points, the instances where the response of the filter is above some threshold in both the images are back-projected to compute the 3D position of these high likelihood body parts. The posteriors for these part locations are then computed by integrating the likelihood with the priors. Both the search methods have been used in our implementation: the likelihood-based approach for the four end-limbs and the prior-based approach for other parts.

One can also use "helpers" to obtain a rough localization of certain parts and to initialize the search process. The most discriminative of these parts is perhaps the face, which may be detected using a face detector (we use a popular one based on [20]). We apply epipolar constraints and matching across views in order to obtain a few helpers in 3D that are used to initiate search in certain high probability regions. Using these "helpers" allows the system to run automatically and efficiently.

The cameras are placed in a wide-baseline configuration to reduce occlusions. The system is able to find parts even if they are visible in only one view and yields a good probability distribution of part location even when the part is not visible in other views. This is due to the inclusion of visibility constraints in the likelihood calculations.

The system flow is shown in Figure 4. The helpers are

[2] although this is not entirely true since the observation is correlated to the apperance of the part that occludes this part.

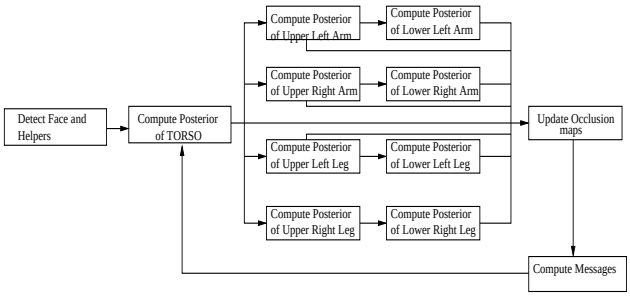
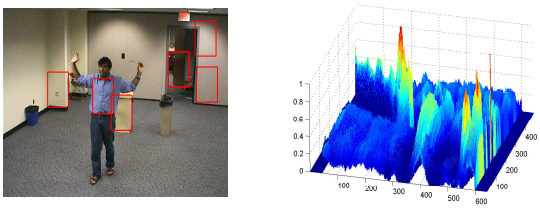


Figure 4. System Flowchart



(a) The Image with likelihood peaks marked

(b) The Image Likelihood

Figure 5. The parallel line feature is very weak as too many parallel lines occur in nature. There is a need of prior based search for fast and accurate detection of body-parts.

first detected using independent detectors. Then, at each iteration, the first step is to find the torso and then search for the other connected parts in turn. The two search methods described above are used to search for each part. Once the posterior distribution of all the parts is estimated at the end of an iteration, messages are passed that update the visibility variables and apply the appearance constraints across parts. The process is iterated until the variance of most of the parts falls below a given threshold.

6. Extension to Tracking in Videos

During tracking, additional temporal consistency constraints can be utilized for more accurate and efficient inference. A simple way to incorporate temporal consistency constraints is to utilize the locations and appearances of different parts at time $t-1$ in order to create priors for locations and appearances of parts at time t. One can incorporate such constraints in a belief propagation framework by adding the potentials $\omega_{t-1,t}(\Phi_i^t, \Phi_i^{t-1})$. The belief propagation equations then change to:

$$\begin{aligned} m_{ij}^r(\Phi_j^t) &= \int \varphi_{ij}(l_i^t, l_j^t)\phi_{ij}(\psi_i^t, \psi_j^t) \\ &\quad P_i(I_1....I_M|L^t)\omega_{t-1,t}(\Phi_i^t, \Phi_i^{t-1}) \\ &\quad \prod_{k\in E_1\setminus j} m_{ki}^{r-1}(\Phi_i^t) \prod_{o\in E_2\setminus j} m_{oi}^{r-1}(\Phi_i^t) \end{aligned}$$

$$\begin{aligned} Pos^r(\Phi_i^t) &\propto P_i(I_1....I_M|L^t)\omega_{t-1,t}(\Phi_i^t, \Phi_i^{t-1}) \\ &\quad \prod_{k\in E_1\setminus j} m_{ki}^{r}(\Phi_i^t) \prod_{o\in E_2\setminus j} m_{oi}^{r}(\Phi_i^t) \end{aligned}$$

$\omega_{t-1,t}(\Phi_i^t, \Phi_i^{t-1})$ can be modeled as product of two gaussians given by

$$\begin{aligned} \omega_{t-1,t}(\Phi_i^t, \Phi_i^{t-1}) &= N(d(l_i^t, l_i^{t-1}), 0, \sigma_{ij}^1) \\ &\quad *N(D(\psi_i^t, \psi_i^{t-1}), 0, \sigma_{ij}^2) \end{aligned}$$

The equations above impose the constraint that one does not expect major changes in location and appearance of a part between two consecutive frames. Also, instead of using full visibility to initialize the belief propagation iterations, we use the occlusion maps estimated from the previous frames. Imposition of such constraints speeds up the inference substantially in tracking applications while also providing temporal consistency.

7. Experimental Results and Evaluation

The anthropometric data for different people was acquired using hand-labeled images. This anthropometric data includes ratios of heights and widths of different body-parts. This data is used for pruning the search area for each body-part. The angular constraints used on body parts were based on the possible movement of the parts. For example, the maximum possible motion between upper arm and lower arm was kept at 150 degrees (assuming the same volume in 3d cannot be occupied by 2 parts). In order to reduce false negatives, the constraints obtained from the hand-labeled images were further relaxed.

We tested the effectiveness of our likelihood model when the parts are visible. Figure 5 shows the computed likelihood of the torso in the image. The result demonstrates the need to prune the search areas based on priors.

Several experiments were performed to demonstrate the importance of the "new" constraints incorporated in our system. The importance of modeling occlusion is demonstrated in Figure 6. In this example, the right upper arm is occluded in view 2 and the right leg is occluded in view 1. Figure 6(a) shows the results of the algorithm in view 1, when occlusion information is not passed and only kinematic constraints are used to find the pose parameters. This would be same as using the algorithm in [18], but using our likelihood model. The results show that when we do not use occlusion information the right leg is totally missed by the algorithm due to confusion with another location. Figures 6 (b) and (c) show the results of the algorithm with all the constraints and occlusion reasoning. When occlusion infor-

(a) (b)

Figure 7. (a) The lower right hand is missed when appearance constraints are not used. b) Appearance consistency with the other hand helps in peaking the posterior at correct location.

mation is passed between the body-parts, the left leg creates a region of occlusion which causes an increase in the likelihood of the right leg being present at its actual location.

In an another experiment, the algorithm was tested without using appearance constraints while occlusion information and kinematic constraints were still used. It can be seen from Figure 7(a) that the lower right arm was missed due to conflicting likelihoods. However, when the appearance constraints are added, correct detection of the lower left arm guides the search for lower right arm as the appearance of the two are expected to be similar [Figure 7(b))].

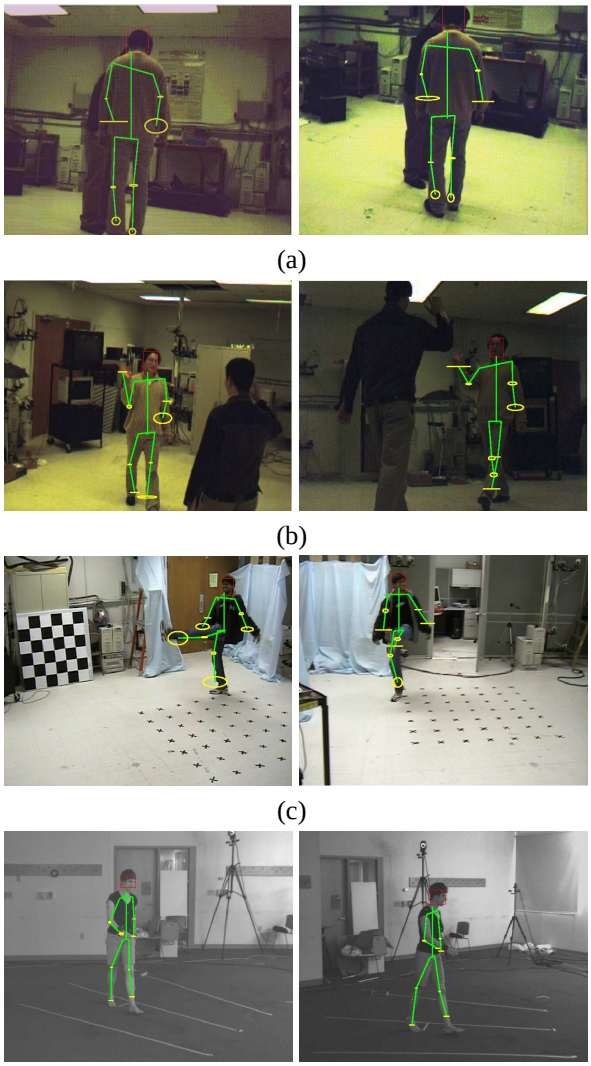
The algorithm was also evaluated when multiple people are present and very close to each other. In such cases, it would be very difficult to first segment one person from the image and hence conventional approaches fail. Figure 8(a) and (b) show the performance of the algorithm in such cases.

Additional results are shown in Figure 8 (c-d); Figure 8(d) is a frame from a commonly used sequence from Brown University [18] .

A few frames of a tracking sequence are shown in Figure 9 (see accompanying videos).

8. Conclusion

We describe an algorithm for estimating the 3D pose of articulated structures such as humans. Probabilistic distribution of various parts are used to compute region of occlusions and compute the probability of visibility of each object part given its location. Unlike previous approaches, where the image likelihoods are computed using the assumption that each part is visible in the image, we compute the image likelihood considering the visibility of parts in different views. We also consider the high correlation between the appearance of left-right part pairs and use it to better localize the part locations. Experimental results demonstrate the importance and effectiveness of incorporating these additional constraints in real scenes with multiple people.



(a) (b) (c) (d)

Figure 8. (a) and (b)Results from a multiple people sequence where background subtraction cannot be used. (c) An unconventional pose (d)Results from a commonly used sequence from Brown University.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, 2004. 1

[2] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *CVPR*, 2003. 1

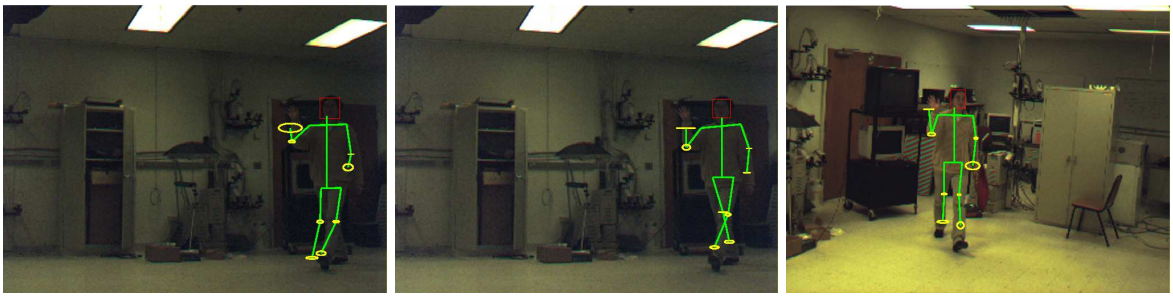
[3] Q. Delamarre and O. Faugeras. 3d articulated models and multi-view tracking with silhouettes. In *ICCV*, 1999. 1

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, volume 2, pages 66–73, 2000. 1

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

[6] G. Hua, M.-H. Yang, and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *CVPR*, 2005. 1

[7] S. Ioffe and D. A. Forsyth. Human tracking with mixtures of trees. In *ICCV*, pages 690–695, 2001. 1



(a) Without Occlusion Constraints View 1 (b) With Occlusion Constraints View 1 (c)With Occlusion Constraints View 2

Figure 6. Illustration of the advantage of using occlusion constraints. Note that the right leg is missed if occlusion constraints are not used.

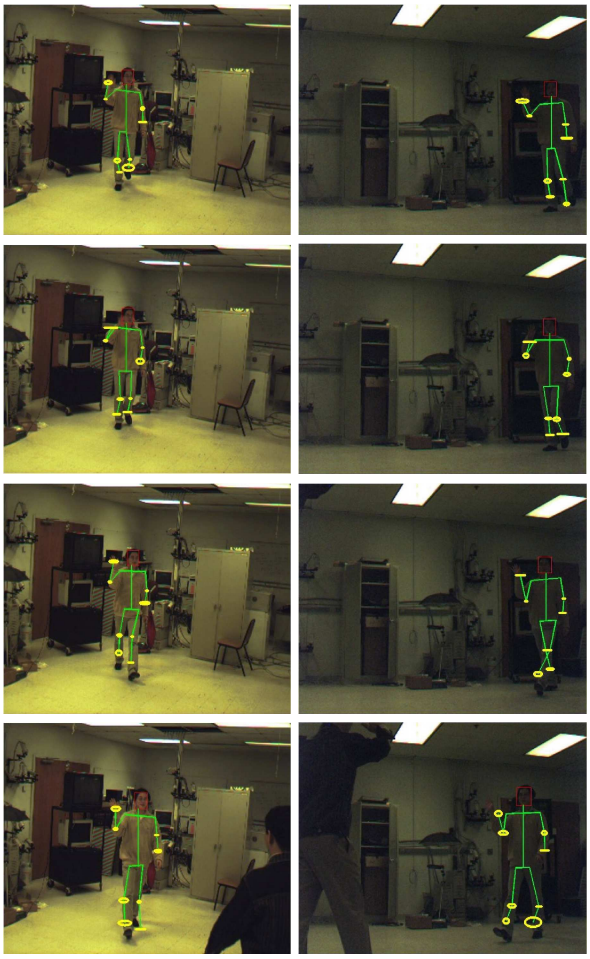


Figure 9. Some tracking results obtained by using temporal constraints. The two columns show the two views of the person.

[8] M. Isard. Real-valued graphical models for computer vision. In *CVPR*, 2003. 3

[9] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005. 1, 2

[10] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *CVPR*, 2004. 1

[11] A. Mittal, L. Zhao, and L. Davis. Human body pose estimation by shape analysis of silhouettes. In *AVSS*, 2003. 1

[12] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, volume 2, pages 326–333, 2004. 1

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

[14] X. Ren, A. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005. 1

[15] S. Roth, L. Sigal, and M. J. Black. Gibbs likelihoods for bayesian tracking. In *CVPR*, 2004. 5

[16] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV*, pages 750–757, 2003. 1

[17] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, volume 2, pages 702–718, 2000. 1

[18] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-linbed people. In *CVPR*, volume 1, pages 421–428, 2004. 1, 2, 3, 7

[19] L. Sigal, M. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *NIPS(16)*, pages 1539–1546, 2004. 1

[20] P. Viola and M. Jones. Rapid object detection using boosted cascade of simple features. In *CVPR*, 2001. 6