

Efficient Human Pose Estimation via Parsing a Tree Structure Based Human Model

Xiaoqin Zhang^{1*}, Changcheng Li^{2*}, Xiaofeng Tong³, Weiming Hu¹, Steve Maybank⁴, Yimin Zhang³

¹National Laboratory of Pattern Recognition, Institute of Automation, Beijing, China

²Department of Automation, Tsinghua University, Beijing, China

³Intel China Research Center, Beijing, China

⁴School of Computer Science and Information Systems, Birkbeck College, London, UK

Abstract

Human pose estimation is the task of determining the states (location, orientation and scale) of each body part. It is important for many vision understanding applications, e.g. visual interactive gaming, immersive virtual reality, content-based image retrieval, etc. However, it remains a challenging task because of unknown image background, presence of clutter, partial occlusion and especially the high dimensional state space (usually 30+ dimensions). In this paper, we contribute to human pose estimation in two aspects. First, we design two efficient Markov Chain dynamics under the data-driven Markov Chain Monte Carlo (DDMCMC) framework to effectively explore the complex solution space. Second, we parse the tree structure state space into a lexicographic order according to the image observations and body topology, and the optimization process is conducted in this order. This realizes a much more efficient exploration than the sampling based search and exhaustive search, and thus achieves a tremendous speed-up. Experimental results demonstrate the efficiency and effectiveness of the proposed method in estimating various kinds of human poses, even with cluttered background, poor illumination or partial self-occlusion.

1. Introduction

Human pose estimation is the task of determining the location, orientation and scale of each body part (i.e. head, torso, upper/lower arms, and upper/lower legs). It is important for many vision understanding applications, e.g. visual interactive gaming, immersive virtual reality, visual surveillance, content-based image retrieval, etc. However, it remains a challenging task because of unknown image background, presence of clutter, partial occlusion and especially the high dimensional state space (usually 30+ dimensions).

*Equal authorship.

Searching for the optimal pose in such a high dimensional state space involves a huge computation time and may get trapped in the curse of dimensionality.

In this paper, we propose an efficient approach for human pose estimation in static images. In our work, human body is modeled as a three-level tree structure and the estimation process is formulated as a Bayesian inference problem. The tree structure state space is carefully parsed into a lexicographic order and solved by the DDMCMC [24] technique. Experimental results demonstrate the efficiency and effectiveness of the proposed approach in estimating various kinds of human poses, even with cluttered background, poor illumination or partial self-occlusion. In addition, our algorithm does not need face or skin detection so that there are fewer constraints on the body pose and the clothing. The main contributions of our approach are:

- We design two promising Markov Chain dynamics under the DDMCMC framework: diffusion and jump, which respectively correspond to a local searching operation and a switch to a new local optimization process. These two dynamics enable us to explore the complex solution space more efficiently.
- The tree structure based human model is carefully parsed into a set of ordered body parts according to the image observations and body topology, and the optimization is conducted in this order. This mechanism realizes a much more efficient exploration than sampling based search and exhaustive search. More importantly, it provides a general mechanism to incorporate prior knowledge (body topology, physical motion) into the optimization process.

The paper is organized as follows. Section 2 introduces related work. The overview of our approach is given in Section 3. Section 4 presents the detailed form of the posterior probability calculation. The estimation process is described

in Section 5. Section 6 shows the experimental results and analysis, and Section 7 is devoted to the conclusion.

2. Related Work

Human pose estimation from static images is an active and popular research area. There are a large number of papers [1, 10, 14, 19, 21] addressing this problem as well as related topics over the past several years. They can be generally divided into the following four categories.

- **Exemplar-based approaches:** in this category, labeled exemplars are stored and used to match to the test image, and pose is assumed to be the same as the most similar exemplar. In [7], a test shape is matched to each stored exemplar in 2D view, using the technique of shape context matching [18] in conjunction with a kinematic chain-based deformation model. Shakhnarovich et al. [9] propose a new algorithm to learn a set of hashing functions which efficiently index examples in the estimation task.

- **Top-down approaches:** methods in this category find promising hypotheses by matching models to the image features. In [4], the estimation is treated as an iterative parsing process, where better features are iteratively built for subsequently parsing. Ramanan et al. [5] track people by detecting them in certain stylized poses repetitively, and discriminative appearance models are learned from a few frames to find persons in latter frames. Sigal et al. [12] propose a variant of the belief propagation method with occlusion reasoning to infer 2D human pose. In [15], the well-known pictorial structures (PS) method is applied to human pose estimation.

- **Bottom-up approaches:** these methods generate part hypotheses by visual cues at the low-level or mid-level, and then extend or assemble them to the whole body by some construction rules. In [8], limbs and torso are detected using a segmentation approach and then assembled into human figures. Srinivasan et al. [16] propose a bottom-up parsing of complete partial body masks guided by a parse tree. At each level of the parsing process, the partial body masks are evaluated directly via shape matching with exemplars. In [17], body parts are first detected and then pieced together using a dynamic programming (DP) procedure. In [22], candidate body parts are detected from bottom-up in parallel, and then these parts are assembled together into body configurations via integer quadratic programming.

- **Top-down and bottom-up approaches:** these methods combine both top-down and bottom-up approaches for pose estimation [2]. In [6], a data driven belief propagation Monte Carlo algorithm utilizing bottom-up visual cues is proposed. Zhang et al. [11] conduct a hybrid strategy combining both deterministic and stochastic search, where visual cues such as edge and skin are used to facilitate the searching process. In [13], the proposal maps generated by

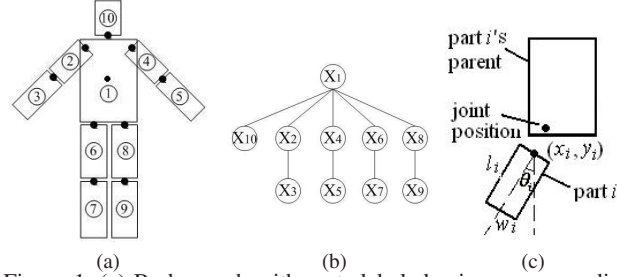


Figure 1. (a) Body graph with parts labeled using corresponding node numbers; (b) tree structure of the body; (c) part states.

observation data are utilized to generate 3D pose candidates during the DDMCMC search.

3. Overview of Our Approach

In this section, we give an overview of our approach containing three major parts: human model, problem formulation and pose estimation framework.

3.1. Human Model

As shown in Fig. 1, the body is modeled as a three-level tree structure, which captures kinematic constraints between body parts. The pose of the body is $\mathcal{X} = \{X_1, X_2, \dots, X_{10}\}$, where X_i represents an individual articulated part i . Each body part is modeled by a rectangle, for which $X_i = \{x_i, y_i, \theta_i, l_i, w_i\}$, where $\{x_i, y_i\}$ denote the 2D location and $\{\theta_i, l_i, w_i\}$ represent orientation, length and width respectively, see Fig. 1.

3.2. Problem Formulation

Denoting I as image observations, pose estimation can be formulated as a Bayesian inference problem for estimating the posterior distribution [20]:

$$P(\mathcal{X}|I) \propto P(I|\mathcal{X})P(\mathcal{X}) \quad (1)$$

where $P(I|\mathcal{X})$ is the likelihood of observations given body states \mathcal{X} , and $P(\mathcal{X})$ is the prior distribution which enforces constraints between body parts. A simple and common solution is the maximum a posteriori (MAP) estimate which is given by

$$\mathcal{X}_{MAP} = \arg \max_{\mathcal{X}} P(\mathcal{X}|I) \quad (2)$$

3.3. Pose Estimation Framework

To give a clear view, the pose estimation framework is schematically shown in Fig. 2. Source image, foreground (obtained by background subtraction [3]) and edge map (obtained by image segmentation [23]) are taken as inputs of the algorithm. In each iteration, the body is first initialized via image observations (see Section 5.1) and body constraints (see Section 4.2), then a local optimization on parsed human body parts under the DDMCMC framework

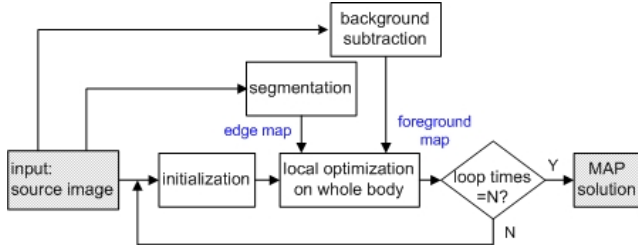


Figure 2. Overview of our approach

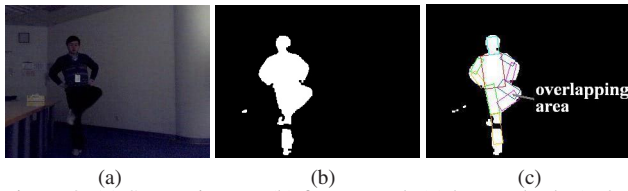


Figure 3. (a) Source image; (b) foreground; (c) human body (color rectangles).

is performed to get a local optimum of posterior probability. The process above is repeated for N times, and the MAP solution is recorded as the pose estimation result.

4. Posterior Probability

In this section, we introduce the details of the posterior probability calculation.

4.1. Observation Likelihood

For the observation likelihood $P(I|\mathcal{X})$ in Eq. (1), we consider both foreground region likelihood $P_f(I|\mathcal{X})$ and edge likelihood $P_e(I|\mathcal{X})$, thus it can be given by

$$P(I|\mathcal{X}) = P_f(I|\mathcal{X})P_e(I|\mathcal{X})^\alpha \quad (3)$$

where α ($\alpha = 0.6$) is a weighting factor for edge likelihood, because it is not as reliable as foreground region likelihood.

4.1.1 Foreground Region Likelihood

Foreground region likelihood measures the degree to which the body agrees with the foreground. For current \mathcal{X} , we synthesize a body and compare it to the foreground, see Fig. 3. Let S_i be the area of the intersection region of the synthesized body and foreground, S_u be the area of the union region of them and S_o be the area of overlapping region among body parts. The foreground region likelihood is formulated by

$$P_f(I|\mathcal{X}) \propto \exp\left(\frac{S_i - \beta S_o}{S_u}\right) \quad (4)$$

where β ($\beta = 0.5$) is a penalty factor for the overlapping part. This likelihood is a maximum when the body agrees with foreground best, i.e. the body covers the foreground as much as possible, and meanwhile covers background and self-overlaps as little as possible.

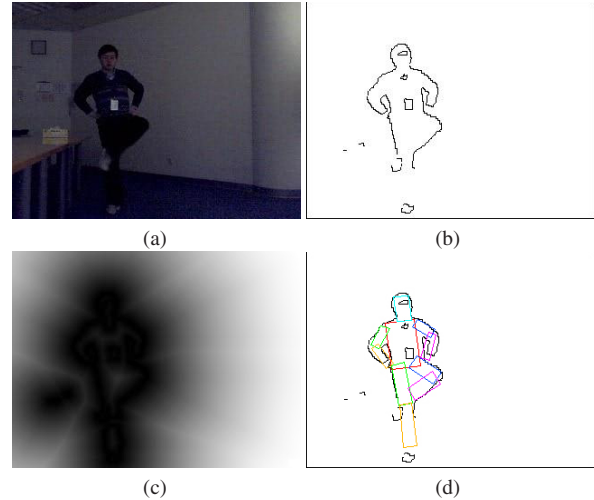


Figure 4. (a) Source image; (b) edge map; (c) edge distance map; (d) human body.

4.1.2 Edge Likelihood

Edge likelihood measures the extent to which the body agrees with image edges. We run image segmentation over the input image and get an edge map (see Fig. 4(b)) by preserving the edges in a body mask generated by dilating the foreground region. After this, we get an edge distance map (see Fig. 4(c)) via distance transform on the edge map. For each point (u, v) on the sides of the synthesized body (color rectangles in Fig. 4(d)), denoting $d(u, v)$ as the value of this point in distance map, E_p as the set of all side points on the synthesized body, and T as the total point number inside E_p . The edge likelihood is then given by

$$P_e(I|\mathcal{X}) \propto \exp\left(-\frac{\sum_{(u,v) \in E_p} d(u,v)}{T}\right) \quad (5)$$

The edge likelihood reaches maximum when the sides of the body agree with image edges best.

4.2. Body Constraints

The prior distribution $P(\mathcal{X})$ in Eq. (1) contains the body constraints, including spatial and length relations among body parts (we choose torso length l_{torso} as the length reference).

As in [15], denoting (x'_i, y'_i) as the joint position for part i in its parent (see Fig. 1(c)), constraints C_i for part i are given by

$$C_i = N(x_i, x'_i, \sigma_{x_i}^2)N(y_i, y'_i, \sigma_{y_i}^2)N(l_i, \mu_l, l_{torso}, \sigma_l^2 l_{torso}^2)N(w_i, \mu_w, l_{torso}, \sigma_w^2 l_{torso}^2)M(\theta_i, \mu_{\theta_i}, m) \quad (6)$$

where $N(\cdot)$ is a Gaussian distribution, and parameters (means and standard deviations) in Eq. (6) are learned from training samples¹. The distribution over the angle θ_i is the

¹For torso, since it has no parent, items about location, orientation in Eq. (6) are set to 1.

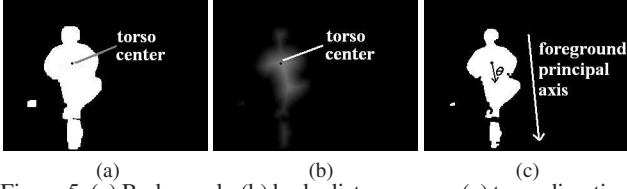


Figure 5. (a) Body mask; (b) body distance map; (c) torso direction and foreground principal axis.

von Mises distribution

$$M(\theta, \mu, m) \propto e^{m \cos(\theta - \mu)} \quad (7)$$

where m is a small constant. Then $P(\mathcal{X})$ can be calculated by

$$P(\mathcal{X}) \propto \exp\left(\sum_i \omega_i C_i\right) \quad (8)$$

where ω_i is the weight of part i which is proportional to the area of part i and fulfills $\sum_i \omega_i = 1$.

5. Estimating Human Pose

5.1. Initialization

Since the body is modeled as a tree structure, we should first initialize the root (torso). The foreground region gives an estimation of body height l_{body} . We assume that $l_{torso} \sim N(\mu_{l_{torso}} l_{body}, \sigma_{l_{torso}}^2 l_{body}^2)$ ($\mu_{l_{torso}}$ and $\sigma_{l_{torso}}$ are learned from training samples), from which we can straightly sample l_{torso} . Based on l_{torso} , w_{torso} can be sampled from the prior distribution mentioned in Section 4.2. For the other states of torso, x_{torso} , y_{torso} and θ_{torso} , we sample them from proposal functions which are given by image observations as follows.

We first run a distance transform on the body mask to get a body distance map as shown in Fig. 5. It is usually a reasonable assumption that the larger a pixel's value is in distance map, the more likely it is to be the center of torso. Based on this assumption, we can obtain the proposal torso center (x_{torso}, y_{torso}) . For θ_{torso} , it is sampled from two parts: perpendicular direction of gradient of torso center in distance map and the direction of foreground principal axis (see Fig. 5).

However, sometimes, the initial torso sampled in this way is bad. Fortunately, it is easy to identify many bad torsos by verifying how the torso agrees with foreground, and this is especially significant to reduce the computation time and increase the accuracy. Here, the criteria for a bad torso are: 1) the area of foreground above shoulders is larger than a certain threshold (see Fig. 6(a)) or 2) the background ratio in head or in torso is larger than a certain threshold (Fig. 6(b)).

After initializing the torso, other parts' states can be sampled from the corresponding prior distributions mentioned in Section 4.2.

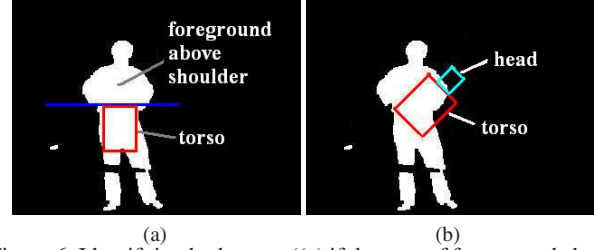


Figure 6. Identifying bad torsos ((a) if the area of foreground above shoulder is large, the initial torso may be upside-down; (b) the torso or head does not match with foreground well).

5.2. Local Optimization

Local optimization is the process of finding a local optimum in the state space of the tree structure. As shown in Fig. 7, there are three main parts in the local optimization process: 1) parsing the state space of tree structure and choosing a body part, 2) choosing a motion type for the chosen body part, 3) deciding whether to accept the new states or not using a Metropolis Hasting approach. Below we give a detailed description of each part.

5.2.1 Parsing the Tree Structure Model and Choosing a Body Part

Traditionally, the tree structure based human model is popularly used to capture the kinematic constraints between human body parts in pose estimation. However, the high dimensional state space employed by the tree structure model is the bottleneck of the human pose estimation. So how to parse the tree structure model into a lexicographic order, and search in a lower dimension is an important issue.

Inspired by the recent progress of tree parsing in the computer languages and natural language processing [25, 26], we propose a tree parsing algorithm for the tree structure based human model according to some *special grammars*. These grammars are defined in the following sections.

5.2.1.1 Grammars for Choosing a Body Part

Our tree structure model is parsed on the following two grammars: image observation and human body topology.

Grammar 1: the body part which agrees with image observation more badly has a priority to be optimized.

Based on *Grammar 1*, we design an importance proposal probability which measures the degree of a body part agreeing with the foreground. In more detail, the human body is synthesized after initialization and is compared with foreground map. As shown in Fig. 8, the following aspects are considered when calculating the importance proposal probability of each body part (see Fig. 8(a)): 1) the area of background in the part region S_{bgIn} ; 2) nearby foreground area not covered by the body S_{fgOut} (we only consider the region inside the dashed rectangle as shown in Fig. 8(a));

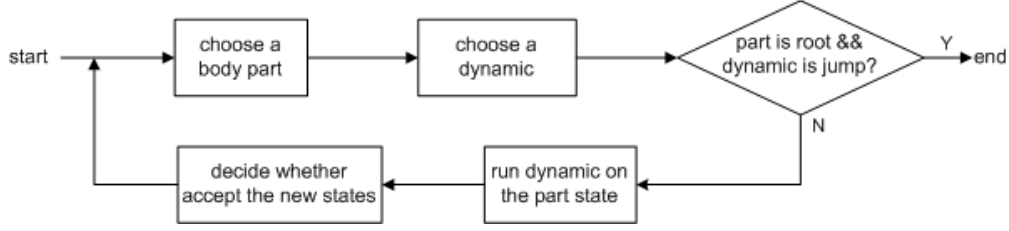


Figure 7. The flowchart of local optimization

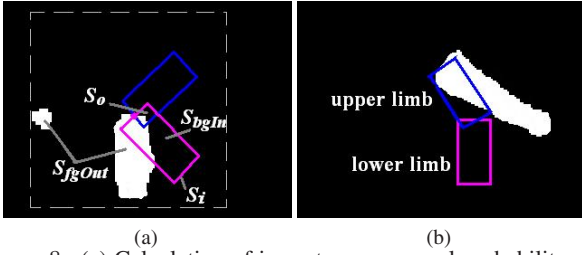


Figure 8. (a) Calculation of importance proposal probability; (b) an example violating the importance proposal probability

3) the area of overlapping region with other parts S_o . Then part i 's importance proposal probability $P_{impt}(i)$ is calculated by

$$P_{impt}(i) \propto \frac{S_{bgIn}}{S_i} + w \frac{S_{fgOut}(S_i + S_o)}{S_i^2} \quad (9)$$

where S_i is the area of part i , and w is a weight. This formula means that a part with large area of background or overlapping region within it, or a large area of uncovered foreground nearby has a large importance proposal probability.

However, merely considering image observations is not sufficient. Fig. 8(b) gives an example, where the upper limb should be optimized before the lower limb even though the lower limb has a larger importance proposal probability. This is because the children node (lower limb) is controlled by its parent node (upper limb), so the parent node is more important than its children.

Grammar 2: the initial importance for each body part decreases from root node to the leaf node, and all the nodes on the same level have the same initial importance.

Therefore, we carefully design a level importance $P_{Limpt}(i)$ for each body part. As a result, the priority for the body part i is formulated as follows,

$$P_{priority}(i) \propto P_{impt}(i)P_{Limpt}(i) \quad (10)$$

From Eq. (10) we can see that image observations and body topology are effectively combined in $P_{priority}(i)$. Therefore, we can parse the tree structure through sequentially sampling the body part based on its $P_{priority}(i)$, and guide the optimization process in this order.

5.2.2 Markov Chain Dynamics: Jump or Diffusion

After choosing a body part, we need to determine the Markov chain dynamic to guide its evolution. In our work,

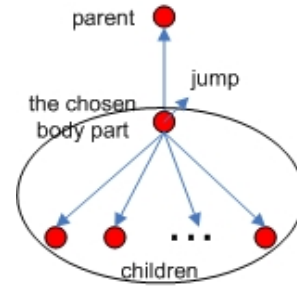


Figure 9. The jump dynamic on the chosen body part activates a new local optimization on the tree structure inside the ellipse.

we design two promising Markov chain dynamics : diffusion and jump, which respectively correspond to a local searching operation and a large change of body part states. The principle for choosing a dynamic for the body part is based on *Grammar 3*.

Grammar 3: If ① the importance proposal probability of the chosen body part is smaller than a predefined threshold or ② the subtree structure (the root of the subtree is the chosen body part)² achieves its local optimum, the jump dynamic is chosen; otherwise, the diffusion dynamic is chosen.

If diffusion is chosen, the state of the chosen body part ϕ (ϕ is one component of $\{x_i, y_i, \theta_i, l_i, w_i\}$) is updated in the following way,

$$\phi' = \phi \pm \lambda \pm \epsilon \quad (11)$$

where λ is step length and ϵ is a Gaussian noise. The sign in front of λ is determined by choosing the one which increases the posterior probability. This updating operation is run repetitively till the posterior probability can not be increased or the state exceeds its range. Diffusion is used to find a local optimum for the chosen part state.

If jump is chosen, we first resample the chosen part state according to the corresponding prior distribution in Section 4.2, and then conduct a new local optimization on the subtree of body (the root of the subtree is the chosen part, please see Fig. 9). These dynamics enable a fast searching (diffusion) or help the optimization process jump out of a local maximum (jump).

These two dynamics are complementary and can be well-organized by *Grammar 3*, and thus can explore the complex

²Except the chosen body part is the root node-torso, in this case, the optimization process ends (see Fig. 7).

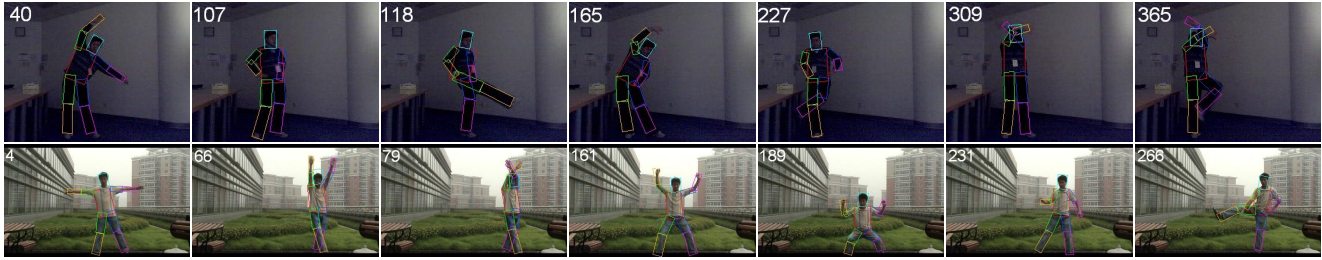


Figure 10. Some results of self-recorded images (first row: indoor, second row: outdoor).



Figure 11. Some results of real world images (first row: Yoga, second row: Shadowboxing).

state space efficiently.

5.2.3 Decide Whether Accept New Pose

The Metropolis Hasting approach is used to determine whether to accept the new state or not. The probability of accepting new pose \mathcal{X}' in place of the current pose \mathcal{X} is given by

$$P(\mathcal{X} \rightarrow \mathcal{X}') = \min \left\{ 1, \frac{P(\mathcal{X}'|I)P(\mathcal{X}|\mathcal{X}')}{P(\mathcal{X}|I)P(\mathcal{X}'|\mathcal{X})} \right\} \quad (12)$$

Here, we assume $P(\mathcal{X}|\mathcal{X}') = P(\mathcal{X}'|\mathcal{X})$ for simplicity, and to avoid the case that bad states are so easy to accept, we add an exponent $k(k > 1)$ for the posterior probability (found empirically from experiments). Then it becomes

$$P(\mathcal{X} \rightarrow \mathcal{X}') = \min \left\{ 1, \frac{P(\mathcal{X}'|I)^k}{P(\mathcal{X}|I)^k} \right\} \quad (13)$$

5.2.4 Level Importance Propagation

Grammar 4: After optimizing the chosen part, the level importance of this part carefully propagates to its parent node and children nodes.

This grammar is based on such knowledge: if a part has just been optimized, its importance should decrease and the decreased value is propagated to other parts so that the other parts become more likely to be chosen for optimizing in next round.

6. Experimental Results

We implement this algorithm with C++ on a platform with Pentium IV 2.8GHz CPU and 512M memory. It is

tested with both self-recorded images and real world images, which cover various kinds of poses, poor illumination and background clutter, etc, and the results are analyzed qualitatively and quantitatively to show the contributions of our algorithm. Some important parameters employed in our work are set as follows: $\alpha = 0.6$, $\beta = 0.5$, $w = 0.2$, $\lambda = 2, 2, 5, 5, 4$ (for the five components of a body part respectively).

6.1. Performance on Self-recorded Images

In this part, we evaluate our algorithm with two sequences of self-recorded images. The first set is captured indoors with very poor illumination, and it contains 380 images with 320×240 size. The second set is recorded outdoors with a cluttered background and it contains 300 images with 400×240 size. Some typical results on the first image set are shown in the first row of Fig. 10. It shows that our algorithm can achieve very good results although the illumination is poor and the poses are rather complex. Furthermore, our algorithm performs well even there are partial occlusions between the body parts (see the frame 118, 227, 309, 365), due to the effectiveness of the kinematic constraints of human body, and more importantly that our searching process is guided by both image observations and body tree topology. The second row of Fig. 10 shows the results of our algorithm on the outdoor image set, from which we can see that our algorithm can handle various poses with a cluttered background. In addition, when the foreground and edge map are noisy, our algorithm can still achieve favorable results, please see section 6.4.³

³We did not use any sequential information in these two image sequences, and all the results were based on the single image.

Methods	Number of iteration	Average running time (by minute)
our algorithm	50	1
Gang's work	6	2-3
Zhang's work	-	5

Table 1. Average running time of our approach and its comparison with Gang's and Zhang's work

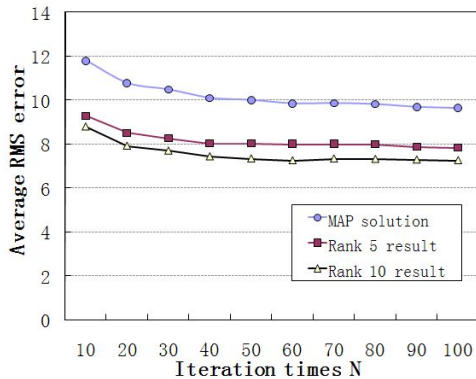


Figure 12. Convergence curves of average RMS errors

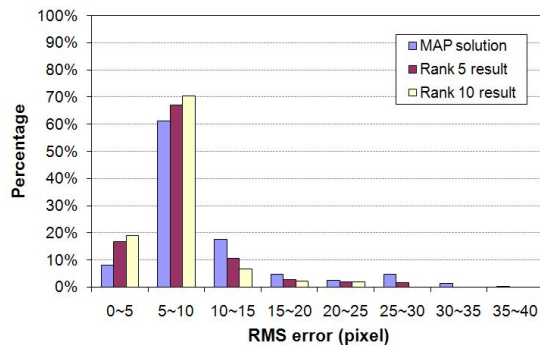


Figure 13. Histogram of RMS errors (N=50)

6.2. Performance on Real World Images

In this part, our algorithm is tested on two real world image sets download from the internet. The first set contains images of a sequence of Yoga action, and in the second set of images, a man is practicing Shadowboxing (a kind of Chinese Kongfu) on a lawn. These two image sets contain many complex and unconventional poses, which are very difficult to estimate. As shown in Fig. 11, our algorithm can give good estimation of body pose even though there are many different and complex actions in these two image sets. More surprisingly, the result is tolerable even when the man turns around and his upper arm is severely occluded (see the second row of Fig. 11). The reason is that the body constraints and the efficient searching enable a correct model fit even when a body part is occluded.

6.3. Quantitative Analysis

To give a quantitative evaluation, we have marked the ground truth by manually locating body joint positions for all the test images. Based on the ground truth, we calculate the Root Mean Square (RMS) errors for each image.

Besides the MAP solution, we also count the RMS errors on Rank 5 and Rank 10 results (Rank 5/Rank 10 result is the one with the lowest RMS error among the top 5/top 10 highest posterior probability results). Fig. 12 shows the convergence curves of average RMS error (unit: pixel) on all frames with the iteration time N increasing. We note that the average RMS errors converge fast and tend to stabilize when $N = 50$. Furthermore, we calculate a histogram of RMS errors for MAP, Rank 5 and Rank 10 results when $N = 50$. As illustrated in Fig. 13, the horizontal axis is RMS error, the vertical axis is percentage of corresponding error occurrences. It shows that about 70% percent of estimation errors fall into the interval of 0-10 pixels for MAP result (for Rank 10 result, about 90%). Since the average RMS of MAP, Rank 5, Rank 10 results are 10.02, 8.04, 7.34 respectively, the histogram shows the stability of our algorithm in the various kinds of images. Meanwhile, to validate the efficiency of our method, we conduct a quantitative evaluation comparison with Gang's work [6] and Zhang's work [11] on average running time. Table 1 shows the results of quantitative comparison. We can see that without any optimization, our algorithm takes about 1 minute to get the result with 50 iterations, which is more efficient than the methods in [6, 11]. In addition, we also conduct a comparison experiment of average running time between our method and the Monte Carlo sampling based searching method without body parsing process. The result demonstrates that the former one can achieve more than ten times speed-up than the latter one.

6.4. Performance With Noisy Inputs

Below, we will give some examples and analysis when the input foreground and edge map are noisy.

As shown in Fig. 14, our algorithm can still get good results when the input edge map is not accurate. This is because we add a control weighting factor when calculating the posterior probability to alleviate the influence of segmentation, see Section 4. Besides, when the foreground extracted is noisy, we may still get good results due to the following reasons (Fig. 14 gives some examples of this case):

1. We dilate the foreground to get the body mask, so that small holes and gaps can be filled;
2. We design two powerful dynamics: diffusion helps to adjust the states to the right values nearby; jump helps the states skip over local minima (foreground holes and gaps).



Figure 14. Some estimation results on noisy foreground or edge map (first row: results, second row: associated foregrounds or edge maps)

7. Conclusion

In this paper, we have proposed a novel approach for human pose estimation based on the DDMCMC framework. With the help of an importance proposal function, level priority of each body part and two well designed Markov chain dynamics, the proposed tree structure parsing algorithm can explore the complex solution space more efficiently and converge to a good result fast. Experiments demonstrate the efficiency and effectiveness of the proposed algorithm on estimating various kinds of human poses, even under cluttered background or poor illumination.

More importantly, the tree structure parsing algorithm is a general framework, which enables to incorporate knowledge (body topology, physical motion) into the optimization process to guide the searching directions, and thus can achieve a tremendous speed-up and much higher efficiency.

Acknowledgment

Main part of this work is accomplished in Intel China Research Center. This work is partly supported by NSFC (Grant No. 60825204, 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453, 2009AA01Z318).

References

- [1] A. Mittal, L. Zhao and L. Davis, "Human body pose estimation using silhouette shape analysis," *AVSS*, 2003. 2
- [2] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference," *CVPR*, 2006. 2
- [3] C. Stauffer, W. Grimson, "Adaptive Background Mixture Models for Real-time Tracking", *CVPR*, 1999. 2
- [4] D. Ramanan, "Learning to Parse Images of Articulated Bodies," *NIPS*, 2006 2
- [5] D. Ramanan, D. Forsyth, and A. Zisserman, "Strike a Pose: Tracking People by Finding Stylized Poses," *CVPR*, 2005. 2
- [6] G. Hua, M. Yang, Y. Wu, "Learning to Estimate Human Pose with Data Driven Belief Propagation," *CVPR*, 2005. 2, 7
- [7] G. Mori and J. Malik, "Estimating Human Body Configurations Using Shape Context Matching," *ECCV*, 2002. 2
- [8] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering Human Body Configurations: Combining Segmentation and Recognition," *CVPR*, 2004. 2
- [9] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast Pose Estimation with Parameter Sensitive Hashing," *ICCV*, 2003. 2
- [10] J. C. Niebles and L. Fei-Fei, "A hierarchical model model of shape and appearance for human action classification," *CVPR*, 2007. 2
- [11] J. Zhang, J. Luo, R. Collins, and Y. Liu, "Body Localization in Still Images Using Hierarchical Models and Hybrid Search," *CVPR*, 2006. 2, 7
- [12] L. Sigal and M. J. Black, "Measure Locally, Reason Globally: Occlusion-Sensitive Articulated Pose Estimation," *CVPR*, 2006. 2
- [13] M. W. Lee and I. Cohen, "A Model-Based Approach for Estimating Human 3D Poses in Static Images," *PAMI*, 2006. 2
- [14] N. R. Howe, M. E. Leventon, W. T. Freeman, "Bayesian Reconstruction of 3D Human Motion from Single-Camera Video," *NIPS*, 1999. 2
- [15] P. F. Felzenszwalb, D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *IJCV*, 2005 2, 3
- [16] P. Srinivasan, J. Shi, "Bottom-up Recognition and Parsing of the Human Body", *CVPR*, 2007. 2
- [17] R. Ronfard, C. Schmid, and B. Triggs, "Learning to Parse Pictures of People," *ECCV*, 2002. 2
- [18] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *PAMI*, 2002. 2
- [19] T. J. Roberts, S. J. McKenna, and I. W. Ricketts, "Human Pose Estimation Using Partial Configurations and Probabilistic Regions," *IJCV*, 2007. 2
- [20] T. Zhao, R. Nevatia, "Bayesian Human Segmentation in Crowded Situations," *CVPR*, 2003. 2
- [21] X. Lan and D. P. Huttenlocher, "Beyond Trees: Common Factor Models for 2D Human Pose Recovery," *ICCV*, 2005. 2
- [22] X. Ren, A. C. Berg, and J. Malik, "Recovering Human Body Configurations Using Pairwise Constraints between Parts," *ICCV*, 2005. 2
- [23] Y. Deng, and B.S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *PAMI*, 2001. 2
- [24] Z. Tu and S. C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *PAMI*, 2002. 1
- [25] Compiler Tools Group, "Tree Parsing", *Technical Report*, University of Colorado, 2002. 4
- [26] M. Collins, "Three generative, lexicalised models for statistical parsing", *In Proceedings of the Annual Meeting of the ACL*, 16-23, 1997. 4