Body Part Detection for Human Pose Estimation and Tracking

Mun Wai Lee and Ram Nevatia
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA 90089-0273, USA
MLee@objectvideo.com , nevatia@usc.edu

Abstract

Accurate 3-D human body pose tracking from a monocular video stream is important for a number of applications. We describe a novel hierarchical approach for tracking human pose that uses edge-based features during the coarse stage and later other features for global optimization. At first, humans are detected by motion and tracked by fitting an ellipse in the image. Then, body components are found using edge features and used to estimate the 2D positions of the body joints accurately. This helps to bootstrap the estimation of 3D pose using a sampling-based search method in the last stage. We present experiment results with sequences of different realistic scenes to illustrate the performance of the method.

1. Introduction

Accurate estimation and tracking of human 3-D body pose from a video stream is important for many applications, especially those requiring understanding of human activities such as for surveillance and video indexing. In these applications, often, only a single camera is available. Also, we do not require or expect the people to be wearing special clothing or be instrumented with markers as is common in the motion-capture applications. We include situations where multiple people are present with possible mutual occlusions.

Pose estimation and tracking from a single video requires solution to several difficult computer vision problems. We must segment humans from the *background* and from each other. Then, we need to find the body parts and estimate their 3-D positions and orientations. Detection of humans is difficult as the image appearance changes with the viewpoint, illumination and clothing. It is hard to separate body parts from each other (and from those of other humans) as many are likely to have similar color and texture (e.g. the arms and the body are usually covered with similar clothing). We must also make 3-D position estimates from a single image. Besides the usual difficulties of estimating 3-D from a single view, several limb positions are ambiguous in a typical view and must be inferred by continuity from other frames.

We assume that the humans to be tracked exhibit some motion and hence they can be separated from the background by use of conventional background modeling techniques. This, however, may cause multiple humans to be merged into a single motion blob. Our task is to separate the multiple humans, find their body parts, make 3-D inferences about their positions and orientations and to track them in an image sequence. A generative approach to pose estimation can be used to search through the pose space for the pose that gives the predicted image that fits best with the observed foreground; such methods have been tried, for example in [3][7] [11][13][14]. While some good results are shown, these methods have high computational complexity as the search space has 30 or more dimensions. Furthermore, some of these systems require good initialization (some by hand) and have difficulty in recovering from tracking failures.

We believe that good bottom-up detection of body parts can help overcome the difficulties of the generative approach. Our goal is not a complete bottom-up process, but rather to seed the generative process to be more efficient and capable of automatic initialization and re-initialization in cases of tracking failures. Our approach is mostly related to previous work that explores good 2D features for pose estimation [2][5][8][9][10][16]. This work is also motivated by findings from other previous work [3][6][7][13] which suggest that global analysis is essential to provide accurate hypotheses evaluation, especially in realistic scenes with self-occlusion and significant background clutter. More recently, data driven belief propagation [4][12] has been proposed for human pose estimation; in contrast to these, our work focuses on reducing computational cost by finding good body part candidates more efficiently.

Detection of body components, such as the torso and the limbs from a single image is difficult due to image variations caused by the varying shape, viewpoint,

IEEE Workshop on Motion and Video Computing (WMVC'07)
0-7695-2793-0/07 $20.00 © 2007 IEEE

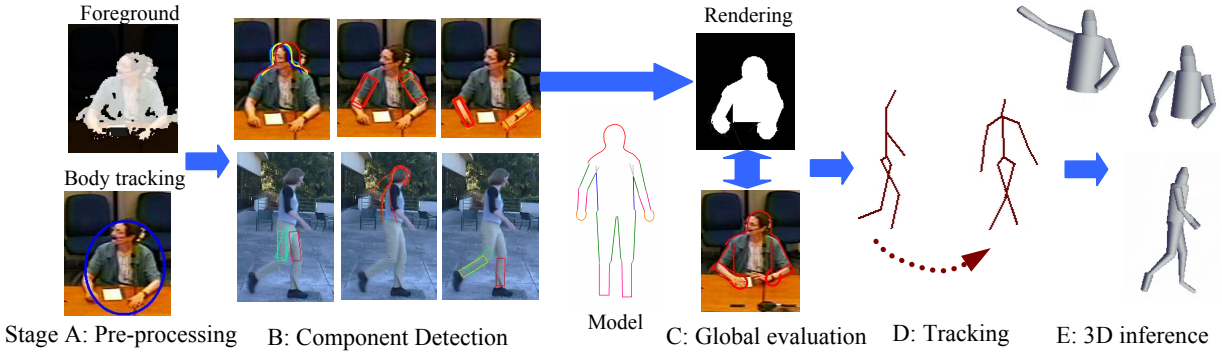


Figure 1: Overview flowchart

clothing and degree of occlusion. We present a new approach and show results on recovery of both the 2D and 3D poses. We require resolution to be high enough so that a person's height is about 150 pixels or more.

2. Approach

An overview of our method is shown in Figure 1. A hierarchical approach is used to extract a set of feature corresponding to the outline of the body and its body components. These features, as represented in the body model shown in Figure 2, are chosen because they are discriminative and are directly related to observable image features. We describe a principled way of extracting, weighing and pruning these features. First a foreground blob is extracted by background subtraction. These blobs are tracked as ellipses representing either full body or upper body (Stage A, Figure 1). Our tracking method incorporates appearance learning to handle inter-occlusions between people. Robust tracking of these ellipses provides coarse estimates of the body position, size and orientation. Features of the individual body components are then detected in a hierarchical approach (Stage B). For each component, multiple candidates are generated and an optimization process is used to align these candidates to image feature accurately. These candidates are then evaluated by an approximate local posterior probability distribution, and candidates with low confidences are removed.

The candidates for different components are combined to form a set of pose hypotheses (Stage C). Each hypothesis is evaluated by computing a global measure of goodness based on a joint posterior probability distribution computed from predictions of the edges of the body. This accounts for the effect of self-occlusion as well as foreshortening. Using multiple pose hypotheses for each frame, an optimal pose trajectory of the sequence is extracted using dynamic programming in Stage D. In Stage E, we introduce additional physical constraints of a 3D articulated human model to estimate the 3D pose using a sampling-based DD-MCMC method.

2.1. Body part-based models

Two part-based models are designed to represent the human body: one for the frontal view and another for the side-view of the person. Each model consists of a set of 12 body components, namely the head, torso, upper and lower arms, hands, and upper and lower legs.

The two models share some similarities as the features pertaining to the limbs are the same. The differences in the two models lie in the features used to represent the upper torso and the head. As shown in Figure 2, the frontal model has a head-shoulder contour (shown in red), while the side-view has a head-shoulder-back contour. Beside this difference, both models have parallel lines representing the limbs. Limbs are essentially cylindrical in shape and limb boundaries project as nearly parallel lines and therefore exhibit some invariance properties that are exploited during detection. By finding these features, we can determine the positions of the body joints in the image.

The model is designed based on our observation of the salient image features of the human body. The main advantage is that the model directly corresponds to good features that are most easily detected in the image. It uses accurate shape contours to represent the upper torso. The features, when combined, form a complete outline of the body, both external and internal. This enables a more accurate evaluation of the hypothesis compared to other models that rely mainly on rectangular blocks. The prior distribution of the shapes of these body components, such as the relative widths and lengths of the rectangles, were derived from a set of training images. This contributes to the generality of the method and the robustness in handling varying shapes of the human body. We note that these two models are used to match not just strictly frontal and profile poses; the modeling of shape

deformation allows matching to other orientations such as 45° view.

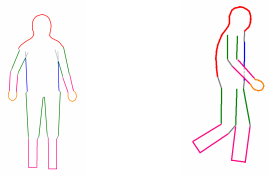


Figure 2: Body outline model. *Left*: frontal. *Right*: side-view

2.2 State estimation framework for 2D pose

We denote $\boldsymbol{\theta}$ as the set of 2D model parameters for a single person; it comprises of $\{\theta_i; I = 1,...,12\}$, where θ_i is a subset of parameters representing one body component, such as the left lower arm. These parameters include position, rotation, and shape parameters.

For an input image I, we want to estimate $\boldsymbol{\theta}$ that maximizes the posterior probability, $p(\boldsymbol{\theta}|I)$ given by the product of prior probability, $p(\boldsymbol{\theta})$, and the likelihood probability, $p(I|\boldsymbol{\theta})$,, i.e.

$$p(\boldsymbol{\theta} \mid I) \propto p(\boldsymbol{\theta}) p(I \mid \boldsymbol{\theta}). \quad (1)$$

By arranging the body components in a tree structure and assuming independence between non-adjacent nodes, the prior distribution can be decomposed into product of pair-wise adjacent nodes:

$$p(\boldsymbol{\theta}) = p(\theta_0) \prod_{(i,j)\in E} p(\theta_j \mid \theta_i), \quad (2)$$

where E is the set of edges in the tree, $p(\theta_0)$ is the prior distribution of the root node, and $p(\theta_j \mid \theta_i)$ is conditional joint distribution of adjacent nodes. These prior distributions are learned from a set of training images. We note here that the training data, described here and in the rest of the paper, consist of some 3D motion capture data as well as annotated images and video sequences. They are from sources different from the test sequences with significant differences in scene and motion. Hence, the learned prior distributions are relatively weak priors, in contrast to stronger priors, such as for walking used in [5].

2.3. Likelihood function

Give a pose hypothesis, we want to compute a global likelihood measure, $p(I|\boldsymbol{\theta})$ to evaluate the hypothesis accurately. We *do not* assume that the observation of different body components are independent. This is because inter-occlusion between body components will affect the observation of each component. Instead, the computation of likelihood function is based on global features after considering the depth order of the components. The depth order is determined by the hypothesized pose.

The likelihood function uses two types of features, edges and foreground, and because they are of different modalities, we assume that the two observations are independent. The likelihood function can be expressed as:

$$p(I \mid \boldsymbol{\theta}) = L_{edge}(\boldsymbol{\theta}) \times L_{fg}(\boldsymbol{\theta}), \quad (3)$$

where $L_{edge}(\boldsymbol{\theta})$ is the global likelihood measure based on edges and $L_{fg}(\boldsymbol{\theta})$ is the likelihood measure based on foreground matching. In the following, we describe these two measures in more details.

The global edge likelihood function is based on Chamfer distance given by:

$$d_{chamfer}(U_{edge}, V_{edge}) = \frac{1}{n} \sum_{u_i \in U_{edge}} \min_{v_j \in V_{edge}} \left\| u_i - v_j \right\|, \quad (4)$$

where U_{edge} is the set of edges in the predicted model after finding the body component (described later); V_{edge} is the set of edges in the input image; $\left\| u_i - v_j \right\|$ is the image distance between two edge points; and n is the number of edge points in U_{edge}. (See Figure 3(a)-(b))

The edge likelihood is given by a one-sided zero-mean Gaussian distribution:

$$L_{edge}(\boldsymbol{\theta}) = \frac{2}{\sqrt{2\pi\sigma^2_{chamfer}}} \exp\left\{ \frac{-d^2_{chamfer}}{2\sigma^2_{chamfer}} \right\}, \quad (5)$$

where $\sigma^2_{chamfer}$ is the variance. This model parameter, as well as others introduced in this paper, is determined empirically using training data according to its definition and the same value is used for all test images. The Chamfer distance can be computed efficiently using distance transform.

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

Figure 3: Global matching using predicted outline and edges. (a) pose hypothesis and predicted edges, (b) extracted input edges, (c) predicted region, (d) extracted foreground (e) error pixels in foreground matching.

Give a pose hypothesis, we predicted the human region as shown in Figure 3(c) and match this with the extracted input foreground as shown in Figure 3(d). The likelihood function is expressed as:

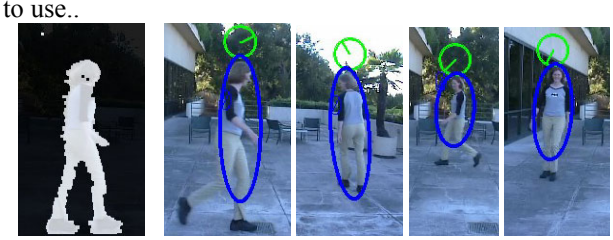
$$L_{fg} = \frac{2}{\sqrt{2\pi\sigma^2_{fg}}} \exp\left\{ \frac{-d^2_{fg}}{2\sigma^2_{fg}} \right\}, \quad (6)$$

where d_{fg} is the number of mismatched pixels (Figure 3(e)), including false alarm and missed detected pixels; and the σ_{fg}^2 is the variance.

3. Body tracking and orientation estimation

The first stage of our approach (Stage A) is the tracking of the entire human body approximated as an ellipse in the image. The ellipse has five parameters: x-y positions, width, height and rotation. The tracking is performed using a cost function is based on region-matching of the estimated ellipses with the extracted foreground blobs (see Figure 4). A color histogram is used to represent the appearance of a human blob and is learned by adaptive updating. When two ellipses overlap, we determine the depth order by comparing the overlapped region with the learned color histograms; the better matched elipse is assumed to be in front.

The direction of motion and the size change of the tracked ellipse are used to give a coarse estimate of the orientation of the person. In Figure 4(b), a green circle is displayed showing the estimated orientation. This estimate is used to determine which of the body models to use..



(a) (b)

Figure 4: (a) Foreground extraction. (b) Ellipse tracking of walking person, with orientation estimation.

4. Hierarchical body component search

A hierarchical approach is used to search for good 2D body components candidates which are later combined for global evaluation. For each component, multiple candidates are generated starting from the face, which is the most discriminative feature, and ending at the lower limbs. At each stage, the result of the previous stage is used to guide the search for the current stage. Each stage involves the following processes labeled as B1 to B4:

B1. Initial positions of the candidate are generated from image information and/or the result of the previous stages. About 50 to 100 initial candidates are generated.

B2. These candidates undergo an optimization process during which their position, orientation and shape are adjusted to better align the candidates to the input image based on edge and foreground cues.

B3. These candidates are weighted by an approximate local posterior probability measure, described later.

B4. Candidates with low weights are removed, with about 10 or less candidates remaining for each body component.

4.1 Head detection

Our head detector uses multiple cues to generate weighted candidates of head position. These candidates are first detected by one of two methods: (*i*) AbaBoost classifier for face detection, and (*ii*) detection of blobs of skin and hair colors using color histogram models. A candidate's weight, $w(\theta_{Face})$, is an approximation of the posterior probability measure of the face given by:

$$w(\theta_{Face}) \approx P(\theta_{Face} \mid \theta_{ellipse}, I) \propto p(\theta_{Face} \mid \theta_{ellipse}) p(I \mid \theta_{Face}), \quad (7)$$

where $\theta_{ellipse}$ are the parameters of the body ellipse and $p(\theta_{Face} \mid \theta_{ellipse})$ is the conditional prior distribution of the face parameters. This conditional distribution is modeled as a mixture of Gaussians.

The likelihood measure $p(I \mid \theta_{Face})$ is given by:

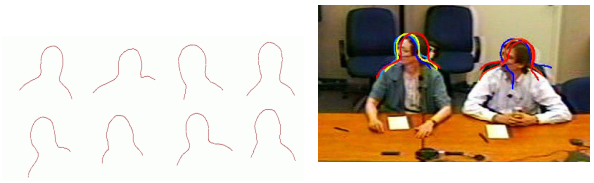
$$p(I \mid \theta_{Face}) \approx L_{Detector}(\theta_{Face}) \times L_{Fg}(\theta_{Face}). \quad (8)$$

The expression for the foreground likelihood for the face, $L_{Fg}(\theta_{Face})$, is similar to Eqn (6), but we only consider false detection of foreground pixels and ignore missed pixels. The term $L_{Detector}(\theta_{Face})$ is the weight of the face detection, which depends on the type of detector that generates the candidate:

$$L_{Detector}(\theta_{Face}) = \begin{cases} P_{TP,AdaBoost} & \text{for AdaBoost candidates} \\ P_{Skin/Hair} \times P_{TP,Skin/Hair} & \text{for skin/hair - ellipse} \end{cases}, \quad (9)$$

where $P_{TP,AdaBoost}$ and $P_{TP,Skin/Hair}$ are the *true positive rates* of the two detectors, and $P_{Skin/Hair}$ is the probability of skin/hair-color of the region, computed by a histogram-based skin-color model. By using multiple cues, the detector has a high detection rate and an acceptable false alarm rate for subsequent processing.

Figure 5: Face candidates from multiple cues including face pattern and skin/hair color region.



(a) (b)

Figure 6: (a) Head and shoulder contour model. (b) Contour detection results. Multiple hypotheses are shown with color coding based on confidences, with red color denoting the highest confidence and blue the lowest.

(a) (b)

Figure 7: (a) Head-back contour model. (b) Detection

4.2 Head-Shoulder or Head-Back Contour

The ellipse tracking stage provides an estimate of the body orientation. For frontal and back-facing poses, we align a head-shoulder contour to the upper body using a gradient descent method. The contour is represented by a set of connected points and the distribution is modeled by a mixture model shown in Figure 6(a).

For initialization, the face detection described earlier is used to determine the starting position of the head hypothesis. The ellipse tracker also provides an estimate of the body size that is used to initialize the scale of head-shoulder contour. For each frame, we extract edges using the Canny detector and apply distance transform on the edge image to derive an energy map. The edge image is first filtered by the detected foreground mask. For optimization (process B2), we perform a gradient descent to find a set of transformations (rotation-scale-translation) for each shape model, so that the contour lies along the regions of low energy. This generates multiple candidates, θ_{HS}. A candidate is weighted by

$$w(\theta_{HS}) \approx p(\theta_{HS} \mid \theta_{Face}, I) \propto p(\theta_{HS} \mid \theta_{Face}) p(I \mid \theta_{HS}) \quad (10)$$

The likelihood is based on edge and foreground features:

$$p(I \mid \theta_{HS}) \approx L_{Edge}(\theta_{HS}) \times L_{fg}(\theta_{HS}). \quad (11)$$

The edge likelihood measure $L_{Edge}(\theta_{HS})$ has the similar expression as Eqn (5), but with a different variance value. For foreground likelihood, we use an expression similar to Eqn (6) :

$$L_{fg}(\theta_{HS}) = \frac{2}{\sqrt{2\pi\sigma^2_{fg,HS}}} \exp\left\{\frac{-d^2_{fg,HS}}{2\sigma^2_{fg,HS}}\right\}, \quad (12)$$

where $d_{fg,HS}$ is the number of false detected foreground pixels, i.e. the number of pixels in the predicted head-shoulder blob that does not lie on the extracted foreground; $\sigma^2_{fg,HS}$ is the variance.

For a side-view pose, where the outlines of one or both of the shoulders are not observed, a *head-back contour* is used in the same manner as the head-shoulder contour. We use a set of 16 contours shown in Figure 7(a), with the head facing right. For a person facing the left, these models are flipped horizontally. Examples of head-back contour detection are shown in Figure 7(b).

4.3. Body Limbs

For the remaining body components, the analysis is performed in the order of the body hierarchical structure, but also in the order of depth. For example, if we have inferred that a person is facing right, we proceed by estimating, the right arm, right leg, left arm, and lastly left leg. From the alignment of the head-shape or head-back contour, the position of the shoulders can be estimated. Candidates for the elbows are generated using corner detectors. Pair-wise combination shoulder and elbow candidates (shown in Figure 8) are used as initial search position for the upper arms.

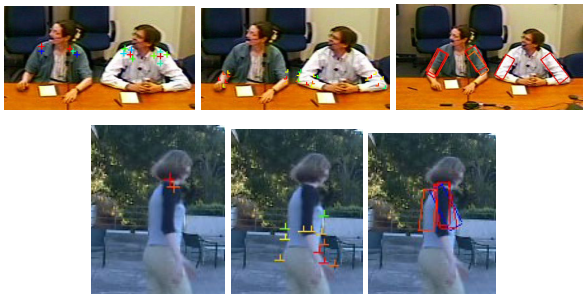
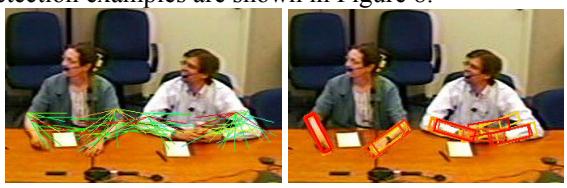


Figure 8: Detection of upper arms. *Left*: Shoulders candidates, and *Center*: elbow candidates are used to initiate search for the upper arms. *Right*: Converged upper arms candidate.

An upper arm is modeled as a rectangle, represented by $\theta_{UpperArm}$, consisting of five degrees of freedom: position (x and y), orientation, length and width; and gradient descent is used to update the rectangle so that it converges to a local maxima of a weight measure, $w(\theta_{UpperArm})$, which is expressed as:

$$w(\theta_{UpperArm}) \approx p(\theta_{UpperArm} \mid \theta_{HS}, I) \propto p(\theta_{UpperArm} \mid \theta_{HS}) p(I \mid \theta_{UpperArm}), \quad (13)$$

where $p(\theta_{UpperArm} \mid \theta_{HS})$, is the conditional probability distribution of the upper arm position given the head and shoulder contour, and learned empirically, and the likelihood $p(I \mid \theta_{UpperArm})$ uses edge and foreground matching and has a similar expression as Eqn (11). Detection examples are shown in Figure 8.



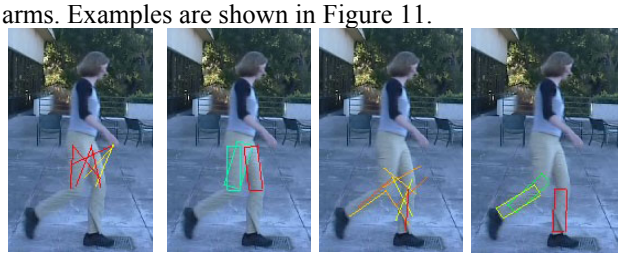
(a) (b)

Figure 9: Detection of lower arms. (a) Lines indicating initial search positions (b) Converged and pruned lower arm candidates.

Figure 10: Detection of Hands (multiple hypotheses are shown)

The detection of the upper arm is then used to guide the search for the lower arm. From the estimated elbow position, multiple search directions are found based on color and foreground, as shown in Figure 9(a), and from these, lower arm rectangles are initialized and converged, via gradient descent, to local maximums (Figure 9(b)). The weight measure $w(\theta_{LowerArm})$ is similar to Eqn (13). Similar approach is used to find the hand, which is modeled as an elliptical contour (see Figure 10).

For estimating full body poses, the tracked body ellipse provides a good initial estimate of the height of the person and a coarse estimates of the hip positions. This is used to first find the upper legs and then the lower legs following the same approach that is used to find the lower arms. Examples are shown in Figure 11.



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

Figure 11: Leg detection: (a) upper legs initial search position (b) converged candidates (c) lower legs initial search position (d) converged candidates.

4.4. Combining the candidates

In previous section, we have described how multiple weighted candidates are generated. We denote $\{\theta_i^j; j = 1,...,n_i\}$ as the set of candidates generated for the ith body component, and n_i is the number of candidates. The set of candidates has been pruned according to their confidences, as described by Eqn (10) and (13), by removing candidates with confidences of less than 10 percent compared to the highest confidence. As a result, n_i is small and usually less than 10.

We now want to find the combination of these candidates, $\boldsymbol{\theta}^* = \{\theta_1^{k1},...,\theta_{12}^{k12}\}$ that maximizes the posterior probability given in Eqn (1):

$$\boldsymbol{\theta}^* \equiv \{\theta_1^{k1},...,\theta_{12}^{k12}\} = \arg\max_{\theta_1^{k1},...,\theta_{12}^{k12}} p(\theta_1^{k1},...,\theta_{12}^{k12} \mid I). \quad (14)$$

The number of combinations is large, a stochastic search approach is used to generate different combinations for evaluation. For each body component, the weights of the candidates are normalized and used as selection probabilities during the search. For the ith body component, we draw samples according to $\theta_i \sim w(\theta_i)$.

4.5. Tracking

For a video sequence, the 2D pose estimation is performed for each frame in the sequence and the best 20 pose hypotheses are found. Dynamic programming is then used to find the best trajectories of the person's motion, similar to the approach used in [6]. The temporal conditional distribution used during dynamic programming is based on a dynamic motion model learned from training motion capture data. This is a generic model that is not specific to any particular type of human motion and can accommodate a wide range of motion.

5. Inference of 3D Pose

The estimated 2D pose trajectory is used to bootstrap the estimation of 3D pose. Despite the loss of depth information, 3D inference can be performed by considering additional physical constraints of the human articulated structure, as demonstrated by the studies in [1] [6]. However, the accuracy of the prior 2D inference should be high; otherwise the depth ambiguity is enormously large. In [6], a human pose estimation framework based on data-driven Markov chain Monte Carlo (DD-MCMC)[15] was proposed and it used bottom-up parts detection to generate proposal of 3D pose. However weak features such as corners and ridge filter responses were used for detecting limbs and as a result, a large number of false alarms were generated. The resulting data-driven proposals have high uncertainties and necessitate a long MCMC search, with an average computation time of 8 minutes per frame.

We adapted the framework in [6], but replaced the parts detection modules with the new method. The

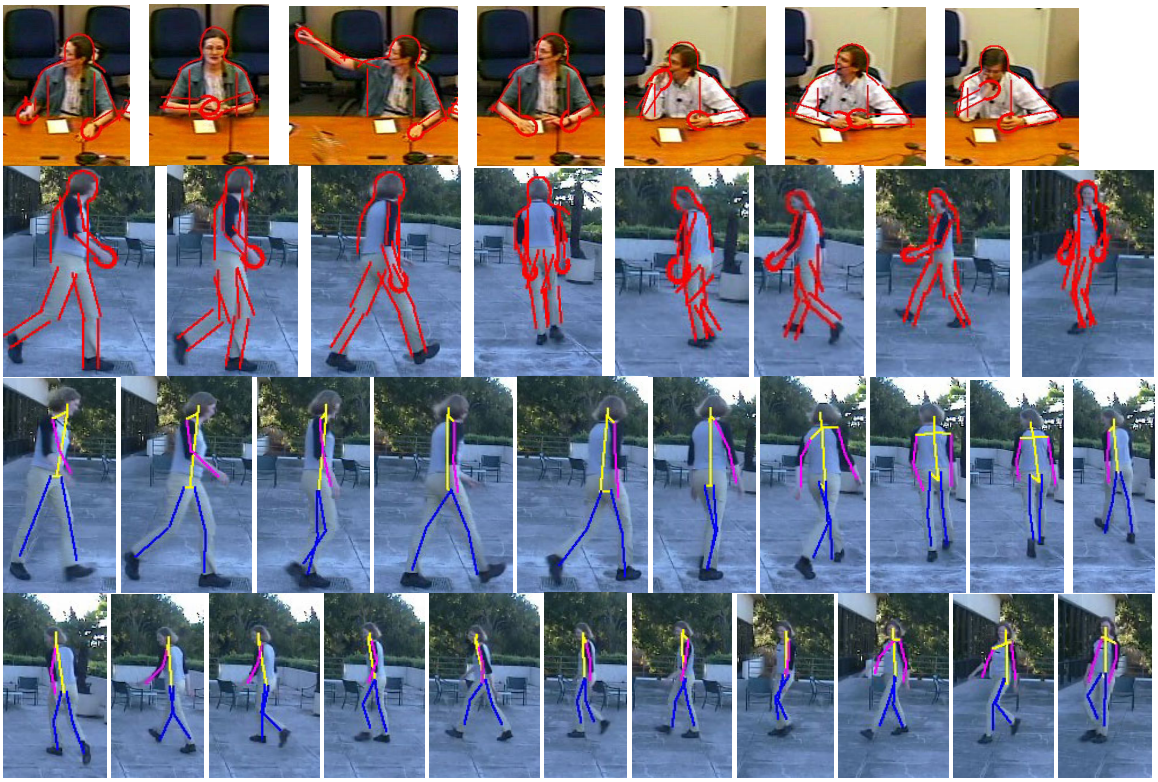


Figure 12: 2D Pose estimation result. *First two rows*: Component detection on two sequences (MAP hypothesis). *Last two row*: 2D poses after tracking (Stage D)

detected pose trajectory is used as bottom-up proposal for the 3D pose estimation, while other proposals, namely random diffusion and depth kinematic jumps, are also used as in [6]. The part detectors provide more accurate proposals and allow the analysis of a broader class of video sequences on which the previous method fails. It also significantly reduces the required computational time.

6. Experiments

For experiments, we use video sequences from different types of realistic scenes, including outdoor surveillance and meeting room video. The type of video is known so that for meeting room video only the upper body poses are estimated. There is no manual initialization of body pose..

Figure 12 shows results of pose estimation, in some selected frames, of a meeting room sequence with 2 persons (400 frames) and in an outdoor scene of a walking person (150 frames) provided by H. Sidenbladh and M. Black. It shows that the outline of the body and limbs are matched with the models accurately. For the outdoor sequence, the system can track the walking person, even though it is not trained specific for this type of motion. The limbs are detected most of the time when they are not occluded. Accuracy of the pose estimation rests heavily on the quality of the candidates of body components in Stage B (described in Section 0). As these candidates are well aligned to the input features, it is easier to distinguish good candidates from the rest. The optimization process in Stage B2, using gradient descent, is therefore important in achieving these good results. At Stage C, there are some errors over the left/right ambiguity of the limbs; a majority of these are resolved during the tracking stage, Stage D.

We manually annotated the joint positions in the test sequences and compute the image distance error of the estimated joint positions. Table 1 shows the average error of each joint. The estimates of head, shoulders and hips are relatively better than the other joints. The overall average error is 9.86 pixels which approximately represent 7 to 12 cm, depending on the resolution of the test sequence.

Joint	Ave. 2D Error	Joint	Ave. 2D Error
Head	9.28	Hips	7.06
Shoulders	9.09	Knees	8.34
Elbows	13.48	Ankles	9.70
Hands	9.52		

Table 1. Breakdown of the error in pixels after 2D tracking (Stage D). The overall average error per joint is 9.86 pixels.

Figure 13 shows examples of 3D inferences where the relatively depths of body joints are estimated using constraints of the 3D articulated structure of the body.

6.1. Computation

Table 2 shows the breakdown of computation time of various processes on an Intel-Xeon 2.8 GHz computer. The average time is 14 seconds per frame for 2D pose inference only and 34 seconds for complete 3D pose inference, which is a significant improvement over a previous method [6] which requires up to 8 minutes/frame. This improvement is achieved because the image positions of most body joints are estimated accurately in Stage C and Stage D, thereby reducing the search space in Stage E.

Figure 13: 3D Pose Inference

Stage	Sec/f	**Stage**	Sec/f
A: Ellipse Tracking	2	D: Tracking (2D)	< 0.1
B: Component detection	6	E: 3D Pose	20
C: Global evaluation	6	**Total**	**34**

Table 2. Computation time (sec. per frame) for each stage.

7. Discussion

We presented a method to estimate and track the human body pose in monocular sequence based on a novel hierarchical approach. Our method allows inference to be made in a coarse to fine manner and from discriminative body components to more difficult components, and we used a principled way for searching candidates of body components and computing their confidences. Combining these candidates results in good hypotheses of the full body pose with accurate alignment of the body model with the input.

The main strengths of this method include the abilities to handle body of different shapes without manual initialization and to overcome significant background clutter and self-occlusion. Experiment results indicate good estimation of the pose. In comparison, the previous method in [6] could not find the body limbs in most frames in the outdoor test sequence due to the low-resolution. With a hierarchical approach, this method is more robust in finding body parts to handle a broader range of scenes and it significantly reduces the computation time when inferring the 3D pose.

Acknowledgment

This research was funded, in part, by the Advanced Research and Development Activity of the U.S. Government under contract # MDA-904-03-C-1786.

References

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

[2] S. Belongie, J. Malik and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *PAMI*, 24(4):509-522, April 2002.

[3] J. Deutscher, A. Davison, I. Reid. "Automatic partitioning of high dimensional search spaces associated with articulated body motion capture," *CVPR 2001*.

[4] G. Hua, M. Yang, Y. Wu, "Learning to Estimate Human Pose with Data Driven Belief Propagation," *CVPR* 2005.

[5] X. Lan, D. P. Huttenlocher, "Beyond Trees: Common-Factor Models for 2D Human Pose Recovery," *ICCV* 2005.

[6] M. Lee, R. Nevatia, "Dynamic Human Pose Estimation using Markov chain Monte Carlo Approach," *Motion 2005*.

[7] M. Lee, R. Nevatia, "Human Pose Tracking Using Multi-Level Structured Models," *ECCV 2006.*

[8] G. Mori, X. Ren, A. Efros, J. Malik, "Recovering Human Body Configurations: Combining Segmentation and Recognition", *CVPR 2004*.

[9] D. Ramanan, D. A. Forsyth, Andrew Zisserman, "Strike a Pose: Tracking People by Finding Stylized Poses," *CVPR 2005*

[10] T. J. Roberts, S. J. McKenna, I. W. Ricketts: "Human Pose Estimation Using Learnt Probabilistic Region Similarities and Partial Configurations." *ECCV 2004*.

[11] L. Sigal, S. Bhatia, S. Roth, M. J. Black, M. Isard, "Tracking Loose-limbed People," *CVPR 2004*.

[12] L. Sigal, M. J. BlacK, "Measure Locally, Reason Globally:Occlusion-sensitive Articulated Pose Estimation," *CVPR 2006*.

[13] C. Sminchisescu, B. Triggs. "Kinematic Jump Processes for Monocular Human Tracking," *CVPR 2003*.

[14] C. Sminchisescu, A. Jepson. "Variational Mixture Smoothing for Non-Linear Dynamical Systems," *CVPR 2004*.

[15] Z. W. Tu, S.C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *PAMI 24(5)*, pp. 657-672, *2002*.

[16] J. Zhang, R. Collins, Y. Liu, "Representation and Matching of Articulated Shapes," *CVPR 2004.*