Human Action Recognition algorithm based on Minimum Spanning Tree of CPA Models

Yi Ouyang
Zhejiang GongShang University, 310018 Hangzhou, China
Email: oyy@mail.zjgsu.edu.cn

Jianguo Xing
Zhejiang GongShang University, 310018 Hangzhou, China
Corresponding author, Email: jgxing.hz@gmail.com

***Abstract*—Human pose recognition algorithm for monocular video was proposed to model human part parameters using video features combination with 3D motion capture data. Firstly Constant Part Appearance (CPA) Models and three-dimensional motion data projection constraint graph structure was defined. To simplify the reasoning process, a constraint graph of the spanning tree construction algorithm and the balancing algorithm were proposed. Combination with the proposed function mechanism, spanning tree of constraint graph and Metropolis-Hastings method, human motion under monocular video can be tracking and recognition, and inferring the 3D motion parameters. By using data-driven (Markov chain Monte Carlo MCMC) and constrain map, human motion limb recognition algorithm is proposed, and the method can be applied to data-driven online human behavior recognition. Experimental results show that the proposed method can recognize human motion action automatically and accurately in monocular video.**

***Index Terms*—human action, Markov chain, belief propagation, 3D human action estimation .**

I. INTRODUCTION

3D modeling of human motion is the mainly support technology for human-computer interaction, animation design, intelligent detection systems and security analysis application system. For non-linear, complex diversity, the lack of a clear classification structure and other characteristics some reasons, human motion modeling is so complicated. At present, for tracking human motion in video image is divided into two categories: based on the Generative Models) [1,2,3] of the human motion analysis and discriminate methods of analysis [4,5,6,7]. Only by observing the movement of the current state of time is often difficult to determine the category of the movement, in order to reduce the computational complexity, constant observation sequence is based on the assumption of conditional independence, the result does not reflect the dependence of the time series. This paper presents an 3D human motion library based on data-driven Markov chain Monte Carlo MCMC method in monocular video to track human motion, the algorithm's basic idea is, using the human body motion capture data building the basic movement database, and at different perspective projection clustering the human silhouette; with [8], [9] method. Monocular video were detected in the human body, and the body can be accurately split the position of the body; Finally, 3D human motion reasoning appearance model algorithm, using time constraints of the model to track the target and graph-driven MCMC and the combination of basic movements, is applied data-driven online actions recognition.

There is a large body of literature on shape representation and matching, see, for example, a recent review by Zhang and Lu [30]. One line of research has dealt with learning shape models from a set of (closed-contour) training shapes.

Finding human body in images is a difficult task, due to the variability in the appearance of people. This variability may be due to the configuration of a person, the human pose and clothing, and variations in illumination. There are two usual strategies for human pose recognition:

A. For searching over model parameters (kinematic variables, camera parameters, etc.) using a comparison between a predicted view of the object and the image. This problem is often stated as optimization of an objective function, which measures the similarity between the predicted and the actual views. This is usually called the top-down approach.
B. Set image features into increasingly large groups, using the current group as a rough hypothesis about the object identity to select the next grouping activity. This is usually called the bottom-up approach.

Much previous work has focused on high-level reasoning (such as mechanisms of inference) but, in our experience, the low-level image features play a crucial role. Background subtraction can be a powerful low-level cue, but does not always apply. We describe human body and texture model to learn appearance. Our system then tracks by detecting the learned models in each frame. We demonstrate the final tracker on hundreds of thousands of frames of commercial and unscripted video. We find that we can accurately track people from a single view with automatic initialization in front of complex backgrounds.

doi:10.4304/jsw.7.7.1577-1584

II. HUMAN MOTION LEVELS MODEL

Using only monocular video for 3D reconstruction of human motion is the kind of ill-posed problem. Mainly due to monocular video camera lack a lot of spatial information. In order to reconstruct each 3D human pose from frames of video images, the basic movement for the human body through the motion sensor to establish a database of basic movements, we use CMU database [18] combined with VOC image database [17] information.

A. Human Body and Texture Model

The human body model (HBTM) is constructed with the torso and limb, in which the location and motion parameters from the torso direction and angle of rotation between the limb composition, as shown in Figure 1. Through the latent variable, shape parameter describes the relationship between the torso and limbs, combined with the common physical description of the color histogram of human appearance. Human joints points is composed by 14 key parts, and pose represented by 6-dimensional vector G, that is the global body's position and direction of rotation . J represents the angle of rotation between joint points. These parameters be modeled by the prior distribution P (G) and P (J). They can be composed by a set of training images, and assume that the probability distribution of approximately Gaussian distribution, and joint points of non-adjacent location parameters are independent. Skin texture model (ST) is composed by $ST = [C_1, C_2, C_3]^T$ the three parameters, respectively, the hands, face and torso rectangle, the prior distribution P (C) is learned from the histogram of training data

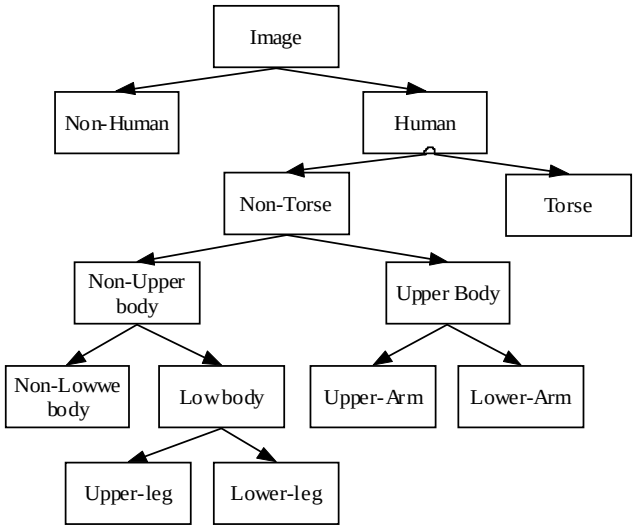

Fig.1 Human parts structure model

B. Cluster-based motion model of the relevant action

As the larger dimension of human motion sequence, and there are large data redundancy, according to the basic model of human body motion data sequence [18] and minimum distance between limbs, we cluster these data at first. This paper presents the Relevant Action Cluster (RPC) for human action analysis. All the input images M select a subset of nodes N in the cluster constraint, which have the largest cluster nodes. Each cluster node will be mapping to a set of images having the same 3D model. This will not only satisfy the action monocular camera image recognition, , simply extend RPC node number of different angles of the projected image, you can improve recognition accuracy having more camera images.

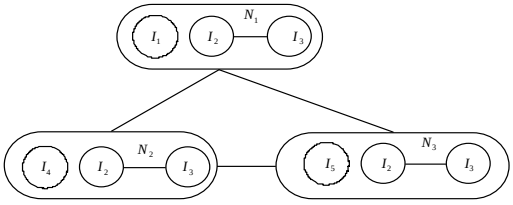

Fig.2 Human motion graph model for RPC nodes

Definition 1: The Relevant Action Cluster (RPC) has $sim(RPC_i, RPC_j) > \varepsilon$,the constraints of the RPC having similar shape, the difference data between the RPC less than 1-ε

Let $I = [I_1, I_2, ...I_M]$ .to represent the image features of human limbs motion. Where $I_i = \{x_k, \theta_j, d, c\}$, M is the number of the RPC nodes, x is the limb center point, and the subscript k means limbs index, superscript j represents the horizontal projection angle index, d as the motion data frame number, c as the basic movement types. 2D projection image feature data is captured from the perspective of the level of 3D motion capture data at different angles. We experiment with 18 different angles, adjacent angles of 10 degrees intervals, and build 2D human motion graph model, where each node corresponds to a cluster RPC node.

Definition 2: RPC graph is G = (V, E, W), where node set V for the RPC, E for the set of edges between v_i, v_j nodes, edges between two nodes which have weight $w_{i,j}$, that is, weighted graph G is undirected, S for the RPC node set size. When at least two nodes RPC images similar, there is a link between two nodes.

III. SPANNING TREE ALGORITHM BASED ON RPC

Using traditional minimum cost spanning tree algorithm, lose some dependencies between nodes, while the uncertainty of tree depth will cause many computing problems, such as the time for determines the reasoning. In order to overcome these problems, we propose a weighted spanning tree construct algorithm based on RPC nodes, node merging algorithm idea is to reduce the size of the spanning tree node, the node splitting to resolve connectivity issues, and make use of spanning tree balancing algorithm remain bounded spanning tree depth.

Definition 3: RPC spanning tree as T = (V, E, W), V as node set, E for the edge set, W is the edge weight set; where each node can only have one parent node, can have many child node.

Node merge: When the same parent node, child node exists between the two sides, when the two sub-nodes associated with edge e intensity threshold intensity is less than Q can be considered an approximate property of the two child nodes, so the two nodes into one new node, its child nodes also point to the node; the two adjacent nodes and edges e deleted. The weight of parent node and the edge of the node is $w = \max(w_i, w_j)$.

Node splitting: When they find a child node also points to more than one parent node, remove the other nodes associated with the strength of the weak side of the parent, to ensure the dependencies between the nodes, will remove even the side of the parent node $N_p^{'}(e)$ connected, that is according to formula (1) with the reservation side of the parent node to be updated.

Spanning tree balance algorithm

Let Child (N), said child nodes of set N, the specific algorithm is as follows:

Step 1. Calculation of the depth of sub-tree and distance vector $Dist(i)$;

Step2. For each sub-tree of length greater than H

From sub-tree node N, the child nodes of N increase the potential edge e ', and weight as

$$w^{'} = w_p * (\mathrm{P(N_i)}/(\sum_{j \in F_i / i} \mathrm{P(N_j)}) * w_c$$

Step3. Consolidate the node Child (N) and the Child (Child (N)).

Step4. Modified sub-tree depth as $Dist(i)$ -1, Goto Step1

For the sub-tree length L is longer than the combined number of sub-root nodes and root node

Assuming the tree depth of 4, when the depth of sub-tree is greater or equal to 4, it will be balanced, as shown in Figure 2. Let N1 for the sub-root node, then the pair node N2, N2, N3 to merge child nodes, while increasing the edge, its weight $w^{'} = \frac{P(N2)}{P(N2+\mathrm{N3})} * 3 * 4$, the combined new node N1 to the N2 and N3 the edge weight as $w_{1,3} = \max(\frac{P(N2)}{P(N2+\mathrm{N3})} * 3 * 4, 4)$, RPC node cluster as Figure 2 and Figure 3.

$$p(X) = \frac{1}{z} p(X_1) \prod_{t=1}^{T-1} p(X_{t+1} \mid X_t) \qquad (4)$$

$$\mathrm{N}_p(e) = \mathrm{N}_p(e) \bigcup (\mathrm{N}_p(e) - N_p^{'}(e)) \qquad (1)$$

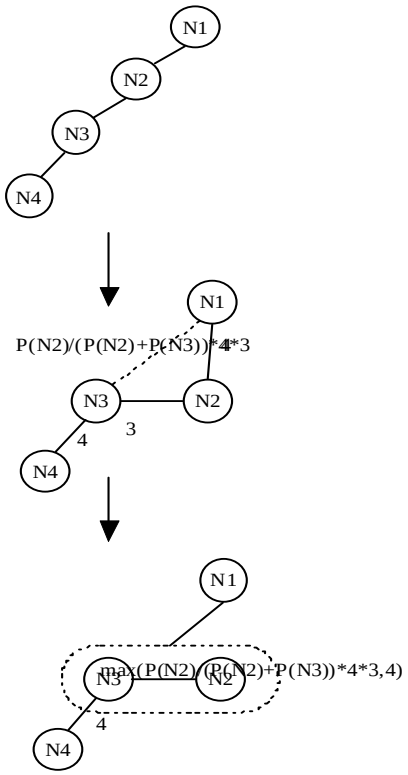

Fig.2 balancing for spanning tree of RPC graph

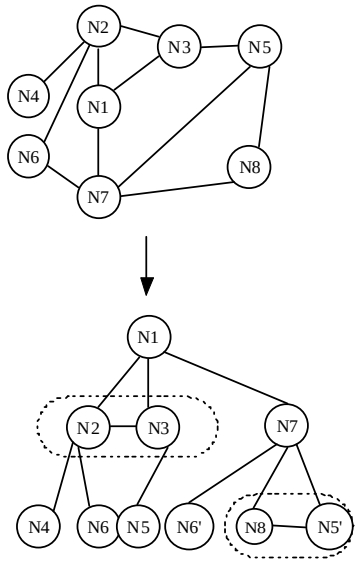

Fig.3 Spanning Tree for RPC graph

IV. HUMAN ACTIONS RECOGNITION ALGORITHM

For human action recognition, such as [3], [11] described the treatment effect directly through the video image is poor, by [3] inspired, we use segmented human motion

recognition technology. We first through the HOG and deformable components of the human body detection method [8], [9] for human motion detection on the first stage; then the spanning tree node with RPC on human action recognition, through this method can be more accurate detection of the human range, and gives the location of the body; Finally, the detection results were sent to the third phase of human space action reasoning.

3D human motion estimation, parameter estimation for the body center is particularly important, it is to consider the perspective of the relationship between operation characteristics and spatial reasoning.

A. The 2D model reasoning based on the RPC

Let $\pi(\cdot)$ as tree node probability, the reasoning process as follow:.

$$\pi(x_v \mid x_{-v}) \propto \pi(x), and\ \pi(x) = \prod_{v \in V} \pi(x_v \mid x_{pa(v)}) \quad (2)$$

$$\pi(x_v \mid x_{-v}) = \pi(x_v \mid x_{pa(v)}) \times \prod_{i, v \in pa(i)} \pi(x_i \mid x_{pa(i)})$$

where V is the set of RPC nodes, pa (i) as the parent node of node i;-v means that all the nodes in V except v.

Prior distribution: Human model for the parameters of each component of the state vector are represented by $X^i_t = \{G,S,C,M\}$ in T frame, where G represents the global position and rotation parameters, S as the parameters that shape, C as that skin color parameters, t denotes the frame number. For simplicity, assuming these parameters independent, prior probability distribution as follow:

$$p(X_t) \propto p(G)p(S)p(C)p(M) \quad (3)$$

These prior probabilities will be combined with image constitutes a posteriori probability function.

B.Constant Part Appearance Models

The approach of clustering part detectors works well when parts are reliable detected. However, building a reliable part detector is hard, a well-known difficulty of bottom-up approaches. An alternative method is to look for an entire person model in single frame. This is difficult because people are hard to detect because of variability in shape, pose, and clothing, just as a well-known difficulty of top-down methods.

We could detect people by restricting our temporal pictorial structure from Equation 4,5 to a single time slice. Such a pictorial-structure detector can cope with pose variation. But since we do not know part appearances C a priori, we can use generic edge templates as part models. Such a detector will be confused by background clutter. We write a single-frame pictorial structure model as:

$$P(part_T, I \mid C_T) \quad (4)$$

$$= \prod_i^N P(part(i) \mid part(pa(i)) P(I(part(i)) \mid part(i), C_i)$$

We use an HBTM model, searching for only one part since we assume the other part will be occluded in our lateral walking pose. We modify the image likelihood terms to look for stylized poses.

The mean shift algorithm finds modes in the log posterior on C given image patches from a sequence $I_t(part_t)$. For tractability, we only consider a finite set of image positions detected by a segment finder. The image likelihoods become linked by the canonical appearance C:

$$P(part_t \mid part_{t-1}) = \Gamma \exp(-\left\| part_t - part_{t-1} \right\|^2) \quad (5)$$

$$P(\mathrm{I}_t \mid part_{t-1}, C) = \Gamma \exp(-\left\| I(part_t) - C \right\|^2) \quad (6)$$

Where Γ is a constant. Our motion model $P(part_t \mid part_{t-1})$ is Gaussian function. Part state $part_t$ encodes both blob position Block(t)=(x,y).Our likelihood model select pixel position Pt at which the encompassing image patch $I(part_t)$ looks like the current blob template C.

We can treat C as a model parameter, then apply the well-known Expectation Maximization algorithm for HMM to infer the hidden variables $part_t$ and the model parameter C. Such an algorithm as follow iterative method:

E-step: Assume we have some estimate positions of the torso appearance C. Then our model can reduces to a standard HMM, and we use dynamic-programming to estimate the sequence of torso positions part(t). One would perform the forward-backward algorithm to compute "soft" positions $P(part_t \mid I_T, C)$

M-step: Given a torso positions from the E-step, we can re-estimate the torso model C by calculating the average image patch at the tracked positions. Just as follow:

$$C' = \frac{1}{T} \sum_t E[I_t(P_t)] \quad (7)$$

$$= \frac{1}{T} \sum_t P(part_t \mid I_T, C) * I_t(part_t)$$

Given the new estimate or torso appearance C', we repeat the E-step. Such a method has a limitations, that is, we do not know the number of people in a given video. If there are multiple people present, then we will have to learn multiple appearance models C.

The resulting model-building algorithm is simple:

Firstly, we select several motion types, such as walking, running etc. and random select one of key frame as $Pos(i)$.

Step1. Initialize a value $C_k = Pos(i)$.

Step2. Find the set of detected patches within a radius of R of C_k.

Step3. If there are many patches from the same frame, only keep the closet to C_k,

Step4. Set C_k to be the average of the set, and if $\|C_{k+1} - C_k\| > \varepsilon$, goto Step 2.

C. The 3D human motion reasoning

The evolution of 3D model of human motion is a known constraint from the start node, by traversing the spanning tree structure, by calculating the maximum a posteriori body image and body movements the best configuration. In order to analyze each type of action, we calculated for each type of action corresponding to the maximum probability, this value was used to measure the movement types of confidence.

Let T frame image sequence of the state vector expressed by $\{ X_1, X_2 ..., X_T \}$. On the shape of human body movement and state of motion relative to the color change more easily. Thus the shape parameters and dynamic parameters should be adjusted so that the object and image in the human motion is consistent, if only to observe the image associated with the current state of the condition. Image prior probability of the state of the human body can be decomposed into a state model with a series of conditions a priori probability of the product:

where Z is the normalization constant. The prior probability sample data can be learned by the training the motion capture data. Conditional probability can be approximated by normal distribution:

Dynamic model in which the covariance matrix, which consists of [18] obtained motion data learning for different types of human movement, such as its value is different. For the state parameters calculated by the probability function. Define the image of the probability function by four parts are: location relationship of limbs; background color difference; human skin color and feature matches. specific probability function as[10] presented the definition.

D. Proposal Function

In [3], [14], [15] were used to top-down [16] method, and [13] method of reasoning to human action, often used for the Metropolis-Hastings MCMC algorithm, the algorithm is the key proposed function of choice, usually by way of random walk, using the proposed function is to generate candidate solutions state. In theory it can produce the entire state sequence of candidate solutions, but more loops in this way. The proposal function will determine the choice of the function convergence speed. This article may be considered state of the current state estimate, the previous state and next state and the model state and the image co-decision.

The previous state X_t^*: Human PRC generated by the current state of the dynamic model estimated parameters, the proposal function is:

$$q(X^*_t \mid X_{t-1}) = q(\mid X^*_t - X_{t-1} \mid) = s \qquad (8)$$

where s is random number between[0,1].

Using of back propagation for spanning tree of the RPC, to obtain another estimate of the current state, after propagation through, making the current state of the estimates take into account future trends.

The final proposal function as follow:

$$\begin{aligned} f_1 &= w_1 \times q(X^*_t \mid X_{t-1}) \\ f_2 &= w_2 \times q(\mathrm{X}^*_t \mid X_t, M^*, I_t) \\ f_3 &= w_3 \times q(\mathrm{X}^*_t \mid X_{t+1}, M^*) \\ q(\mathrm{X}^*_t \mid X_t, M^*, I_t, X_{t-1}, X_{t+1}) &= f_1 + f_2 + f_3 \end{aligned} \qquad (9)$$

Where is the weight factor, and the factor used to adjust the proportion between the various proposed components, experiments were taken 0.3,0.4,0.3.

E. Parameters Learning and Inference

First, we introduce some notation. Binary state random variable X', Bg denotes the presence of an object O or background Bg at a particular node and image location. At the leaf level of the tree, the object class O occurs for the best matching template at the best location. Furthermore, the object class O_l occurs at level l for the optimal path from the root to the best matching leaf level node, together with the associated locations on the coarse-to-fine image grid. The dissimilarity measurement obtained at the lth level of the tree, associated with random variable D, is denoted by

Desired is a Bayesian framework for modeling the a posteriori probability of the object class at a particular node of the tree, given (dissimilarity) measurements along the path to that node.

Inference: We use the information transmission as the main inference. Because E is a tree structure, we first installed from the bottom up approach to news from part (i) delivered to the part (j).

Information is calculated as follows:

$$m_i(l_j) \propto \sum_L \varphi(part(i) - part(j)) a_i(part(i)) \qquad (10)$$

$$a_i(part(i)) = \phi(part(i)) \prod_k m_k(part(i)) \qquad (11)$$

Transmission of information using circular convolution of the way, the response image was first

converted to (Height × width, theta (i)) two-dimensional vector, with the direction of the convolution weights, to derive all directions on the part of the response . Then each part with its child nodes by multiplying the response for the transmission of information, as show in Figure 4.

Each pixel in the image from the root node from torso,head, left arm and left wrist of the probability distribution (as show in Figure 4) (for the accumulation and distribution in all directions

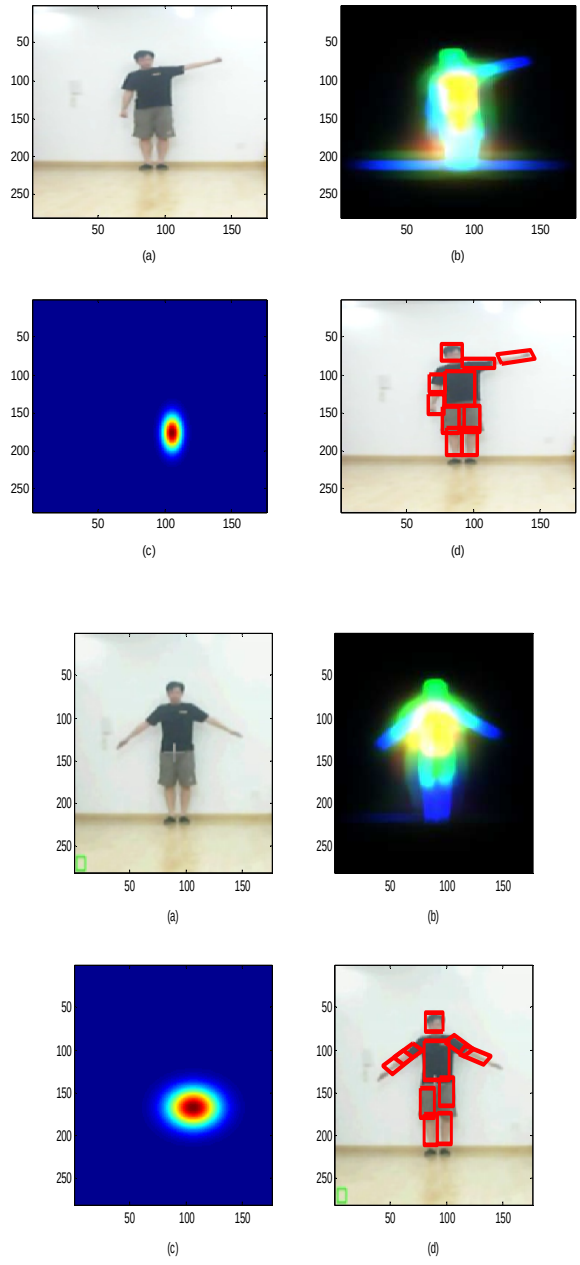

Fig.4 probability distribution in all directions on the part of the response

The second step, bottom-up reasoning:

Dynamic programming to solve the problem parameters: As show in Table 1.

Define the forward variable

Table 1 .Dynamic programming algorithm for image parameters in the human body posture

Step1.Calculate the direction along the edge of the template observation of the image response $TP(part_i, \theta_j)$

Step2. Initial computing

$$\alpha_1(i) = \pi_i TP_i(O_1) \qquad 1 \le t \le T$$

Step3.Use top-down method to determine the human body model

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_i(i) a_{ij}] TP_j(O_{t+1}) \qquad 1 \le t \le T$$

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Setp4. And then use bottom-up approach combined with the rhythm data to identify the body region movement

Step 5 Repeat Step2 until you meet the convergence criteria

V. EXPERIMENT ANALYSIS

More than 200 were collected from a subset of the different motion sequences, each containing 10 sample subset of actions, each action of 18 different angles from the projection, one of the motion sequence in Figure. 2. We test 20 kinds of basic movement, which contains a variety of action walking, running, jumping, kicking, boxing. For different types of data movement, we estimated the depth of Z-axis parameters for the test identification error, A group for general walking, B group running, C group hand boxing. were set before the experiment the depth of the spanning tree when the RPC was 30, this method and Data-MCMC [3] methods and do not use RPC model MCMC method of reasoning directly compare the experimental error of measurement such as Table 2:.

Table 2 Recognition errors of three class motion using different algorithm

Weighted average errors	walking	runnin g	boxing
Non RPC model	25.35	30.18	35.54
Data-MCMC	24.47	27.61	26.53
Our method	21.36	23.15	24.14

The proposed action recognition algorithm based on RPC . We compared it with non-RPC structures methods and ,Data-MCMC method . Our method improved the recognition accuracy. The reason is considered a priori three-dimensional information of human basic movement, the results shown in Table 3.

Table 3 Errors of human parts for walking sequence

Error type	Center	Depth(Z)	Depth(Z)
Torso	16.21	11.2	7.58
RUL	17.24	14.2	5.12
RLL	15.21	12.5	3.12
RSL	15.12	10.32	4.14
RUA	13.52	12.1	6.78
RLA	15.75	13.4	4.34

Fig.5 running 2D human pose estimation results

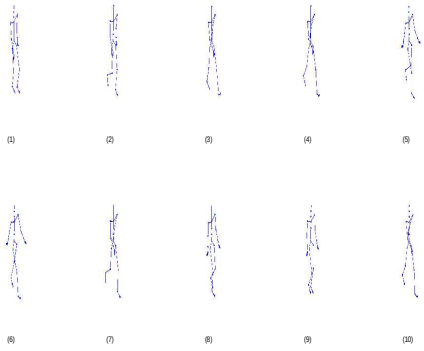

Fig.6 The results of 2D running human pose estimation matching with 3D human motion capture data.

A post processing procedure was used to deal with cases where both arms are visible. First the sampled arm shapes are divided into two clusters based on hand positions, and the mean shape of each cluster is computed. Then we compare the hand distance between these two mean shapes to the width of the torso. If the ratio is above a threshold of 0.6, then both arms are assumed to be detected. To quantitatively evaluate the proposed model, we randomly selected 50 images and hand-labeled the ground truth boundaries of body parts, the 2D human pose estimation results as show in Figure 5. The motion capture data just as Figure 6,7. The results of matching with 3D human motion recognition results as Figure 8.

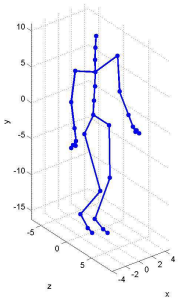

Fig.7 one of 3D Human running pose

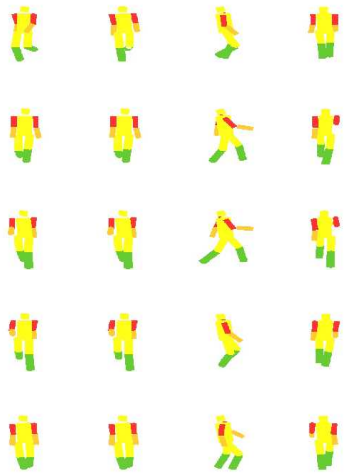

Fig.8 Walking and running pose estimation results

ACKNOWLEDGMENTS

This work is supported by the Science and Technology Department of Zhejiang Province in China No. 2011C24008, No. Y1110809, No.2008C14100 and the Department of Education of Zhejiang Province Project NO Y200907404

REFERENCES

[1] Agarwal, A.; Triggs. Recovering. 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1): 44 - 58 .

[2] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference, EEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006: 1743-1752

[3] LEE, M.W. and R. NEVATIA. Human pose tracking in monocular sequence using multilevel structured models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(1). 27-38.
[4] Weiwei Guo Patras, I. Queen Mary. Discriminative 3D human pose estimation from monocular images via topological preserving hierarchical affinity clustering, Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, 2009: 9 -15.
[5] Ankur Agarwal, Bill Triggs. 3d human pose from silhouettes by relevance vector regression, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), 2004(2):882-888.
[6] A. Elgammal and C.S.Lee. Inferring 3d body pose from silhouettes using activity manifold learning, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04),2004 (2):681-688
[7] Lv F, Nevatia R. Single view human action recognition using key pose matching and viterbi path searching, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04),2007(4): 1−8 .
[8] P. Felzenszwalb, D. McAllester, D. Ramaman. A discriminatively trained, multiscale, deformable part model, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'08),2008:1-8.
[9] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009,32(9):1627-45..
[10] M. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04),2004: 334-341
[11] M. Lee and R. Nevatia. Dynamic human pose estimation using markov chain monte carlo approach motion, IEEE Workshop on Motion and Video Computing, 2005. WACV/MOTIONS'05, 2005(2):168-175.
[12] WANG, Y., N.L. ZHANG, and T. CHEN. Latent tree models and approximate inference in Bayesian networks . Journal of Artificial Intelligence Research, 2008,32(1). 879-900
[13] D. Ramanan, D.A. Forsyth, and A. Zisserman. Strike a pose:tracking people by finding stylized poses, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04),2004 (1):271-278.
[14] Z.W. Tu and S.C. Zhu, Image segmentation by data-driven markov chain monte carlo, IEEE Trans. Pattern Analysis and Machine Intelligence, 2002(24)5:657-672.
[15] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment, Proceedings of. IEEE Conf. Computer Vision and Pattern Recognition, 2004 (2):406-413.
[16] S. Zhu, R. Zhang, and Z. Tu. Integrating bottom-up/top-down for object recognition by data driven markov chain monte carlo, Proceedings of. IEEE Conf. Computer Vision and Pattern Recognition,2000(1): 738-745.
[17] EVERINGHAM, M. AND VAN~GOOL, L. AND WILLIAMS, C. K. I. AND WINN, J. AND ZISSERMAN, EB/OL.2008.A.The PASCAL VOC2008 Results.
[18] CMU,Cmu motion capture library.EB/OL.2007 http:,mocap.cs.cmu.edu.

Yi Ouyang received the MSE degree from College of Computer and Information Engineering, Zhejiang Gongshang University, in 2005. He is currently working toward the PhD degree at the Zhejiang University. He preside over several project of the science and technology of Zhejiang Province, and participating in many National and province natural science funds projects, etc. His research interests are in distributed parallel processing, image processing and machine learning theory and applications.