Chapter 94
Human Action Recognition Algorithm Based on Minimum Spanning Tree

Yi Ouyang and Jianguo Xing

Abstract Human pose tracking recognition algorithm for monocular video was proposed to model human part parameters using video features combination with 3D motion capture data. Firstly three-dimensional 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, 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.

Keywords Human action · Markov chain · Belief propagation · 3D human action estimation

94.1 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

Y. Ouyang (✉) · J. Xing
Zhejiang GongShang University, 310018, Hangzhou, China
e-mail: oyy@mail.zjgsu.edu.cn

X. He et al. (eds.), *Computer, Informatics, Cybernetics and Applications*,
Lecture Notes in Electrical Engineering 107, DOI: 10.1007/978-94-007-1839-5_94,

modeling is so complicated. At present, for tracking human motion in video image is divided into two categories: based on the Generative Models [1–3] of the human motion analysis and discriminate methods of analysis [4–7]. 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 chapter 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.

94.2 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 [17] combined with VOC image database [10] information.

94.2.1 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. 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 six-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

94.2.2 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 chapter 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.

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 $1 - \varepsilon$

Let $I = [I_1, I_2, \ldots I_M]$ 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.

94.3 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 (94.1) with the reservation side of the parent node to be updated.

$$\mathrm{N}_p(e) = \mathrm{N}_p(e) \cup (\mathrm{N}_p(e) - N'_p(e)) \tag{94.1}$$

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 $Dist(i)$;

Step 2 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

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

Step 3 Consolidate the node Child (N) and the Child (Child (N)).

Step 4 Modified sub-tree depth as $Dist(i)-1$, Goto Step 1

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 Fig. 94.1. 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 $w_{1,3} = \max\left(\frac{P(N2)}{P(N2+\mathrm{N3})} * 3 * 4, 4\right)$, RPC node cluster as Fig. 94.1.

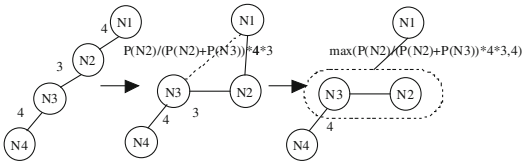

Fig. 94.1 Balancing for spanning tree of RPC graph

94.4 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.

94.4.1 The 2D Model Reasoning Based on the RPC

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

$$\begin{aligned} \pi(x_v|x_{-v}) \propto \pi(x), \quad and \;\; \pi(x) = \prod_{v \in V} \pi(x_v|x_{pa(v)}) \\ \pi(x_v|x_{-v}) = \pi(x_v|x_{pa(v)}) \times \prod_{i, v \in pa(i)} \pi(x_i|x_{pa(i)}) \end{aligned} \tag{94.2}$$

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 $X_t^i = \{\mathrm{G,S,C}, M\}$ 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:

$$p(X_t) \propto p(G)p(S)p(C)p(M) \tag{94.3}$$

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

94.4.2 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 $\{X_1, X_2 \ldots, X_T\}$. 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:

$$p(X) = \frac{1}{z} p(X_1) \prod_{t=1}^{T-1} p(X_{t+1}|X_t) \tag{94.4}$$

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 [12] presented the definition.

94.4.3 Proposal Function

In [3, 13, 14] were used to top-down [15] method, and [16] 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 chapter 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:

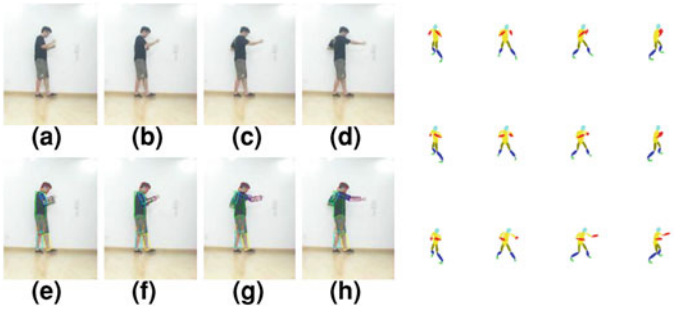

Fig. 94.2 Recognition for running motion sequence

$$q(X_t^*|X_{t-1}) = q(|X_t^* - X_{t-1}|) = s \tag{94.5}$$

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:

$$q(\mathrm{X}_\mathrm{t}^*|X_t, M^*, I_t, X_{t-1}, X_{t+1}) = w_1 \times q(X_t^*|X_{t-1}) + w_2 \times q(\mathrm{X}_\mathrm{t}^*|X_t, M^*, I_t) + w_3 \times q(\mathrm{X}_\mathrm{t}^*|X_{t+1}, M^*) \tag{94.6}$$

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.

94.5 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 Figs. 94.2 and 94.3. 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 94.1:.

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 94.2.

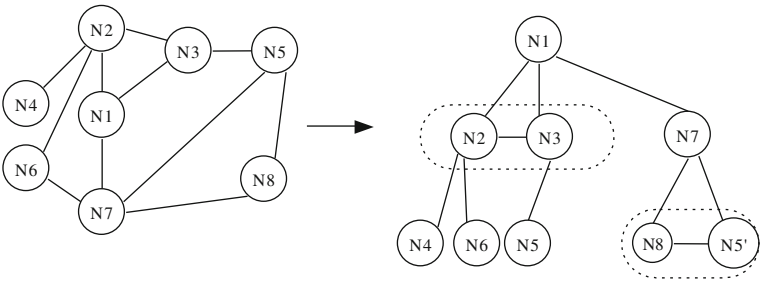

Fig. 94.3 Spanning Tree for RPC graph

Table 94.1 Recognition errors of three class motion using different algorithm

Weighted average errors	Walking	Running	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

Table 94.2 Errors of human parts for walking sequence

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

Acknowledgments This work is supported by the Science and Technology Department of Zhejiang Province in China No. 2008C14100 and Department of Education of Zhejiang Province Project NO 1130KZ710019G

References

1. Agarwal A, Triggs B (2006) Recovering. 3D Human pose from monocular images. IEEE Trans Pattern Anal Mach Intell 28(1):44–58
2. Sminchisescu C, Kanaujia A, Metaxas D (2006) Learning joint top-down and bottom-up processes for 3d visual inference. In: IEEE Computer Society conference on computer vision and pattern recognition (CVPR'06), pp 1743–1752
3. Lee MW, Nevatia R (2008) Human pose tracking in monocular sequence using multilevel structured models. IEEE Trans Pattern Anal Mach Intell 31(1):27–38
4. Weiwei G, Patras I, Queen M (2009) Discriminative 3D human pose estimation from monocular images via topological preserving hierarchical affinity clustering. In: IEEE 12th international conference on Computer vision workshops (ICCV Workshops), 2009, pp 9–15
5. Agarwal A, Triggs B (2004) 3D human pose from silhouettes by relevance vector regression. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 2, pp 882–888
6. Elgammal A, Lee CS (2004) Inferring 3D body pose from silhouettes using activity manifold learning. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 2, pp 681–688

7. Lv F, Nevatia R (2007) Single view human action recognition using key pose matching and viterbi path searching. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 4, pp 1–8
8. Felzenszwalb P, McAllester D, Ramaman D (2008) A discriminatively trained, multiscale, deformable part model. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'08), pp 1–8
9. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2009) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9): 1627–1645
10. Everingham M, Van-Gool L, Williams CKI, Winn J, Zisserman A (2008) EB/OL.2008. The PASCAL VOC 2008 Results. CMU,Cmu motion capture library. EB/OL.2007 http:mocap.cs.cmu.edu
11. Lee M, Nevatia R (2005) Dynamic human pose estimation using arkov chain Monte Carlo approach motion. IEEE workshop motion video Comput 2005. WACV/MOTIONS'05 2, pp 168–175
12. Lee M, Cohen I (2004) Proposal maps driven mcmc for estimating human body pose in static images. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 2, pp 334–341
13. Tu ZW, Zhu SC (2002) Image segmentation by data-driven markov chain monte carlo. IEEE Trans Pattern Anal Mach Intell 24(5):657–672
14. Zhao T, Nevatia R (2004) Tracking multiple humans in crowded environment. Proc IEEE Conf Comput Vis Pattern Recognit 2:406–413
15. Zhu S, Zhang R, Tu Z (2000) Integrating bottom-up/top-down for object recognition by data driven markov chain monte carlo. Proc IEEE Conf Comput Vis Pattern Recognit 1, pp 738–745
16. Ramanan D, Forsyth DA, Zisserman A (2004) Strike a pose: tracking people by finding stylized poses. IEEE Comput Soc Con Comput Vis Pattern Recognit (CVPR'04) 1, pp 271–278
17. Wang Y, Zhang NL, Chen T (2008) Latent tree models and approximate inference in Bayesian networks. J Artif Intell Res 32(1):879–900