Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 506752, 9 pages
http://dx.doi.org/10.1155/2013/506752

Research Article

Adaptive Self-Occlusion Behavior Recognition Based on pLSA

Hong-bin Tu, Li-min Xia, and Lun-zheng Tan

School of Information Science and Engineering, Central South University, ChangSha, HuNan 410075, China

Correspondence should be addressed to Li-min Xia; xlm@mail.csu.edu.cn

Received 31 July 2013; Accepted 24 October 2013

Academic Editor: Feng Gao

Human action recognition is an important area of human action recognition research. Focusing on the problem of self-occlusion in the field of human action recognition, a new adaptive occlusion state behavior recognition approach was presented based on Markov random field and probabilistic Latent Semantic Analysis (pLSA). Firstly, the Markov random field was used to represent the occlusion relationship between human body parts in terms an occlusion state variable by phase space obtained. Then, we proposed a hierarchical area variety model. Finally, we use the topic model of pLSA to recognize the human behavior. Experiments were performed on the KTH, Weizmann, and Humaneva dataset to test and evaluate the proposed method. The compared experiment results showed that what the proposed method can achieve was more effective than the compared methods.

1. Introduction

Automatic recognition of human actions from video is a challenging problem that has attracted the attention of researchers in the recent decades. It has applications in many areas such as entertainment, virtual reality, motion capture, sport training [1], medical biomechanical analysis, ergonomic analysis, human-computer interaction, surveillance and security, environmental control and monitoring, and patient monitoring systems.

Occlusion state recognition has been traditionally tackled by applying statistical prediction and inference methods. Unfortunately, basic numerical methods have proved to be insufficient when dealing with complex occlusion scenarios that present interactions between objects (e.g., occlusions, unions, or separations), modifications of the objects (e.g., deformations), and changes in the scene (e.g., illumination). These events are hard to manage and frequently result in tracking errors, such as track discontinuity, inconsistent track labeling.

The Pictorial structure method [2], which represents the human body as a set of linked rectangular regions, does not take occlusion into account. Sigal et al. [3] argue that the self-occlusion problem can be reduced by an occlusion-sensitive likelihood model. This works well if the occlusionstates (i.e., the depth ordering of parts) is known; for example, if it is specified at the start of the motion and then does not change over time. But, in practice, the depth order of object parts—for example, right arm, torso. Estimating 2D human pose is difficult because of image noises (e.g., illumination and background clutter), self-occlusion, and the varieties of human appearances (i.e., clothing, gender, and body shape) [3–5]. Estimating and tracking 3D human pose is even more challenging because of the large state space of the human body in 3D and our indirect knowledge of 3D depth [6]. In contrast, our approach focuses on self-occlusion. While all of the above methods are modeled to estimate poses from still images, there exists only limited research on the same task in videos. Guo et al. [7] applied the BOW model with human action recognition in video sequence. Niebles et al. [8] successfully applied this model to classify the video sequence of the human action. Wang and Mori [9] assigned each frame of an image sequence to a visual word by analyzing the motion of the person it contains. Sy et al. [10] applied the CRF with a hidden state structure to predict the label of the whole sequence of human gestures. Sigal et al. [3] modeled self-occlusion handling in the PS framework as a set of constraints on the occluded parts, which are extracted after performing background subtraction which renders it unsuitable for dynamic background scenes.

Our work follows literatures [3, 7, 9, 11] by producing a framework for articulated pose estimate-on robust to

cluttered backgrounds and self-occlusion without relying on background subtraction models. The step of rectifying occluded body parts via a GPR model is inspired by recent work by Asthana et al. [12] who used GPR for modeling parametric correspondences between face models of different people. Our problem is more difficult because the human body includes more parameters to be rectified and has more degrees of freedom than faces.

In order to overcome the shortcomings mentioned above, we propose an adaptive self-occlusion state recognition method that estimates not only everybody configuration but also the occlusion states of body parts.

Firstly, the Markov random field was used to represent the occlusion relationship between human body parts in terms of occlusion state variable by phase space obtained. Then, we proposed a hierarchical area variety model. Finally, we infered human behavior by pLSA. Experiments on Human Eva data set were performed to test and evaluate the proposed algorithm. The experiment results have shown that the proposed method is effective in action recognition.

2. Human Trajectory Reconstruction

A tree structure movement of the human body skeleton structure is used by creating visual invariant model [13], the human body is divided into 15 key points; namely, 15 joint point represents the human body structure, and the 15 joints trajectory represents the human body behavior and then uses Markov random field (MRF) by calculating the observation, spatial relations, and the motion relationship and ultimately determines the occlusion positions of the body joints and restores the missing trajectory. Specific steps described below.

The Markov random field (MRF) was used with a state variable representing the occlusion relationship between body parts. Formally, the MRF was a graph $G = (V, E)$, where V was the set of nodes and E was the set of edges. The graph nodes V represented the state of a human body part and graph edges E model the relationships between the parts [11]. The probability distribution over this graph was specified by the set of potentials defined over the set of edges. The MRF structural parameters are defined as follows: $X_i = (x_i, y_i, z_i)$: The ith joint point coordinates; $X = \{X_1, X_2, \ldots, X_{15}\}$: extract the key points of the body 15; $\gamma(X_i)$ $(i \leqslant 15)$: the ith joints visible parts, this parameter is used to determine occlusion relation between nodes. When occlusion occurred, trajectories intersected between

$$X_i\left(X_i\left(x_i, y_i, z_i\right)\right), \qquad X_j\left(X_j\left(x_j, y_j, z_j\right)\right); \tag{1}$$

$\Lambda = \{\Lambda_{ij}\}$ $(i \leqslant 15, j \leqslant 15)$: the occlusion relation among the 15 body joints. When $\Lambda_{i,j} = 0$, the ith and $j$the joints do not occluded. When $\Lambda_{i,j} = 1$, the ith occluded jth. When $\Lambda_{i,j} = -1$, the jth occlude ith; $\lambda_i = \{\lambda_1, \ldots, \lambda_{15}\}$: the ith occlude joints node; then, potential of kinematic relationship is calculated as follows:

$$\psi_{ij}^{K}\left(X_i, X_j\right) = N\left(d\left(x_i, x_j\right); \mu_k, \delta_K\right) f\left(\theta_i, \theta_j\right). \tag{2}$$

This function indicates the position of two adjacent joints, and the angles among joints.

$d(x_i, x_j)$ is the Eucidean distance between two adjacent joints. $N()$ is the normal distribution with $\mu_k = 0$ and standard deviation δ_K.

$E_{O|\wedge}$: occlusion area belong to joints; $W_i = \{w_i\}$: If i joint is occluded, $w_i = 1$, if i joint is not occluded, $w_i = 0$; I: input image; v_{ij} Indicator for overlapping body parts; $\phi_i(I, X_i; \wedge_i)$: potential of observation; $\phi_i^C(I, X_i; \wedge_{ij})$: potential of the color; $\phi_i^E(I, X_i; \wedge_i)$: potential of the edge; $\phi_i^{C_{\text{visible}}}(I, X_i; \wedge_{ij})$: the motion state of X_i (the ith body joint) in the viewing area; $\phi_i^{C_{\text{occluded}}}(I, X_i; \wedge_{ij})$: the motion state of X_i (the ith body joint) in the occluded area; ϕ_i: potential of observation; ψ_{ij}^K: potential of kinematic relationship; ψ_i^T: potential of temporal relationship. Defining a model, similar to [12] for calculating three potential function as follows.

Firstly, we get the observation potential function:

$$\phi_i\left(I, X_i; \wedge_i\right) = \phi_i^C\left(I, X_i; \wedge_i\right) + \phi_i^E\left(I, X_{\text{i}}; \wedge_{\text{i}}\right). \tag{3}$$

The potential of the color

$$\phi_i^C\left(I, X_i; \wedge_{ij}\right) = \phi_i^{C_{\text{visible}}}\left(I, X_i; \wedge_{ij}\right) + \phi_i^{C_{\text{occluded}}}\left(I, X_i; \wedge_{ij}\right), \tag{4}$$

where the first term is X_i of probability of occurrence of color in the visible area and the second term is for the occluded area. The visible term is formulated as

$$\begin{aligned}\phi_i^{C_{\text{visible}}}&\left(I, X_i; \lambda_i\right)\\ &= \prod_{u \in (\gamma(X_i) - (\gamma(X_i) \bigcap \gamma)(X_j))} P_C\left(I_u\right)\\ &= \prod_{u \in (\gamma(X_i) - (\gamma(X_i) \bigcap \gamma)(X_j))} \frac{P\left(I_u \mid \text{foreground}\right)}{P\left(I_u \mid \text{background}\right)},\end{aligned} \tag{5}$$

where $P(I_u \mid \text{foreground})$ and $P(I_u \mid \text{background})$ are the distributions of the color of pixel u given the foreground and background.

$$\begin{aligned}\phi_i^{C_{\text{occluded}}}&\left(I, X_i; \lambda_i\right)\\ &= \prod_{u \in (\gamma(X_i) \bigcap \gamma(X_j))} \left[z_i\left(I_u\right) + \left(1 - z_i\left(I_u\right)\right) P_C\left(I_u\right)\right]\end{aligned} \tag{6}$$

and $z_i(I)$ is calculated as follows:

$$z_i\ (I)\ \frac{1}{N} \sum \left(\phi_j^C\left(I_u, x_j\left(t\right); \lambda_j\right)\right) \tag{7}$$

$u \in (\gamma(X_i) \bigcap \gamma(X_j))$: the occlusion area is determined by the calculated overlapping region of X_i and X_j, N is the sum of all occlusion nodes.

When $f(\theta_i, \theta_j) = 1$, $T_{\text{lower}} \leq \theta_i - \theta_j \leq T_{\text{upper}}$, where T_{lower} and T_{upper} are the lower and upper bound of motion area between X_i and X_j defined by kinesiology.

Finally, potential of temporal relationship is calculated as follows:

$$\psi_i^T\left(X_i^t, X_i^{t-1}\right) = p\left(X_i^t - X_i^{t-1}; \mu_i, \sum_i\right), \tag{8}$$

where μ_i is the dynamics of X_i at the previous time step and $\sum_i$ is a diagonal matrix with a diagonal element is identical to $|\mu_i|$, which similar to a Gaussian distribution with the time.

In this paper, the posterior distribution of model X conditioned on all input images up to the current joint s structure, the current time step τ and occlusion state variable $\wedge^{1:\tau}$ is

$$p\left(X^{\tau} \mid I^{1:\tau}; \wedge^{1:\tau}\right) = \frac{1}{Z} \exp\left\{-\sum_{i \in X^{1:\tau}} \phi_i^C\left(I, X_i; \lambda_i\right) - \sum_{ij \in E_K^{1:\tau}} \psi_{ij}^K\left(X_i, X_j\right) - \sum_{i \in E_T^{1:\tau}, t \in 1:\tau} \psi_i^T\left(X_i^t, X_i^{t-1}\right)\right\}, \tag{9}$$

where Z is a normalization constant.

In a word, we put ϕ_i, ψ_{ij}^K, ψ_i^T into (4), and get body occluded joints positions,

$$\widehat{X}^t = \underset{X^t}{\operatorname{argmax}}\, p\left(X^t \mid I^{1:t}; \widehat{\wedge}^{t-1}\right), \tag{10}$$

where X^t is X joint location at t time.

The occluded relation among joints can be obtained by formula (2).

$$\widehat{\lambda}_{i,j}^t = \underset{\wedge_{ij} \neq 0}{\operatorname{argmax}}\, \phi_i\left(I^t, \widehat{X}_i^t; \wedge_{ij}\right), \tag{11}$$

where $\widehat{X}_i^t$ is X_i position at t time.

The occluded joints can be calculated by MRF at the entire time of motion. In this paper, we connect missing data in order to restore missing coordinate position.

3. Feature Representation

The human action can be recognized in terms of hierarchical area model, relative velocity, and relative acceleration.

3.1. Hierarchical Area Model. For describing the human motion pose (e.g., jogging, running, and walking), we make use of hierarchical area model and extract human facial area S^H, upper limbs area S^U and leg area S^L. To human facial area S^H are extracted in the following way.

(1) According to Canny algorithm, each of the facial contour point set is extracted, and denoted as Ck, where k is the number of contour point.

(2) The face contour can be least square fitting by Ck, which obtained in step 1.

(3) According to step 1 and step 2, if the body movement to make the front, the face area is the largest, if the human turned sideways, the face area will change. Thus, face area in coordinate is

$$\Delta S_x^H(y,z) = \frac{\sum_{i=1}^n S^H\left(x^i, y, z\right)}{\sum_{i=1}^n \delta\left(S^H\left(x^i, y, z\right)\right)},$$
$$\Delta S_y^H(x,z) = \frac{\sum_{i=1}^n S^H\left(x, y^i, z\right)}{\sum_{i=1}^n \delta\left(S^H\left(x, y^i, z\right)\right)}, \tag{12}$$
$$\Delta S_y^H(x,y) = \frac{\sum_{i=1}^n S^H\left(x, y, z^i\right)}{\sum_{i=1}^n \delta\left(S^H\left(x, y, z^i\right)\right)},$$

where n is the frames, $S^H(x^i, y^i, z^i)$ is the set of face contour in all frames, $\delta(v(x^i, y^i, z^i))$ is the set of contour in all frames.

(4) By Repeat Steps 1~3, the face area can be calculated in all frames.

Calculating S^U and S^L is similar to S^H.

Figure 1 shows that the curve for some area features of pedestrian walking. Figure 1(a) is the area variation curve of S^H. Figure 1(b) is the area variation curve of S^U. Figure 1(b) is the area variation curve of S^L.

3.2. Relative Velocity and Relative Acceleration. We can get the relative velocity and relative acceleration by the trajectory of each joint.

Each point' weight can be considered as the same, and build statistical model to calculate the relative velocity and relative acceleration among relative motion joints (e.g., hands and legs) in order to reason the initial state of motion.

$$\Delta_{i,j} = \frac{p\left(x_i(t)_v, x_j(t)_v\right)}{\sum_{k=1}^n p\left(x_i(t)_v, x_j(t)_v\right)}, \tag{13}$$

where $\Delta_{i,j}$ is the relative velocity among i and j.

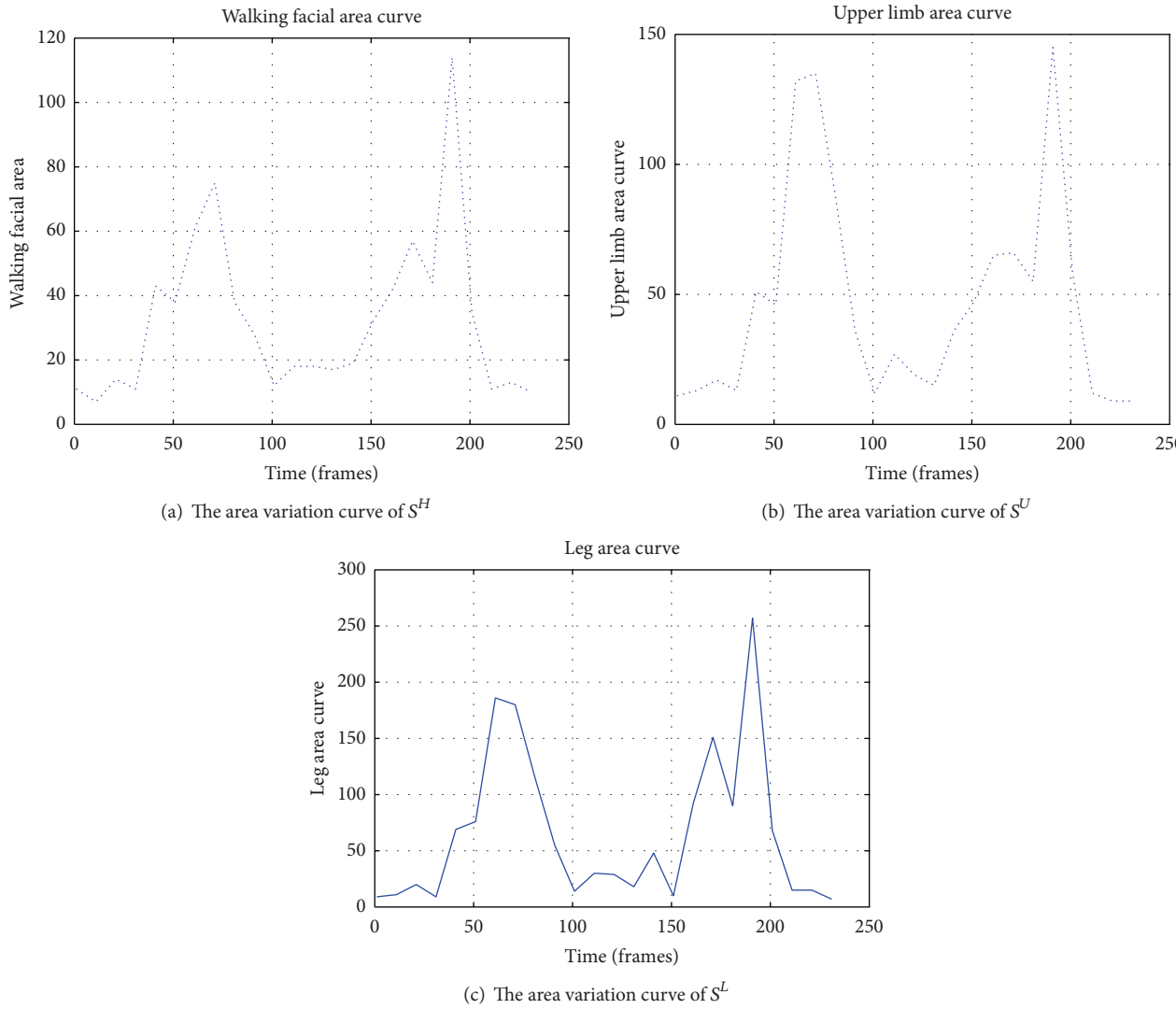
The area-velocity goodness T_j is obtained as follow.

T1: jogging, Δv (the left knee, the right knee), Δv (the left foot, the right foot), Δv (the right knee, the right foot), Δv (the left foot, the left ankle), Δv (the right foot, the right ankle) > $t1$, and $\Delta\alpha$ (the left foot, the left knee) > $t2$.

T2: running, Δv (the left foot, the left knee), Δv (the right foot, the right knee), Δv (the left foot, the left ankle), Δv (the right foot, the right ankle) > $t3$, and $\Delta\alpha$ (the left foot, the left knee), $\Delta\alpha$ (the left foot, the right knee), and $\Delta\alpha$ (the left foot, the right foot) > $t4$.

T3: walking, Δv (the left foot, the left knee), Δv (the right foot, the right knee), Δv (the left foot, the left ankle), and Δv (the right foot, the right ankle) > $t5$.

T4: jumping, Δv (the left foot, the left knee), Δv (the right foot, the right knee), Δv (the left foot, the left ankle), Δv (the right foot, the right ankle) > $t6$, and $\Delta\alpha$ (the left foot, the left ankle), and $\Delta\alpha$ (the right foot, the right ankle) > $t7$.



(a) The area variation curve of S^H

(b) The area variation curve of S^U

(c) The area variation curve of S^L

FIGURE 1: The curve for some area features of pedestrian walking.

T5 : boxing, Δv (the left foot, the left knee), Δv (the right foot, the right knee), Δv (the left foot, the left ankle), Δv (the right foot, the right ankle) $> t8$ and, $\Delta\alpha$ (the left hand, the left elbow), $\Delta\alpha$ (the right hand, the right elbow), $\Delta\alpha$ (the left foot, the left ankle), and $\Delta\alpha$ (the right foot, the right ankle) $> t9$.

Thresholds ($t1$, $t2, \ldots,$ $t9$) are determined empirically as 1.5, 40, 5.5, 60, 3.5, 5.0, 40, 7.0, and 30.

We cluster the extract feature, which meet the threshold requirement, and extract the typical behavior of the action dataset as a standard action: jogging, running, walking, jumping and boxing. Above 5 kinds of common action decomposition, we get relative velocity among joints, when some action occurred. For example, an jogging operation, the relative velocity of the left leg and the right leg and the relative velocity of the left leg and the left knee are more than others joints.

3.3. Codebook Formulation. In order to construct the codebook, we use the k-means algorithm based on the Euclidean distance to cluster all the features (hierarchical area model, relative velocity and relative acceleration) extracted from the training frames. The center of each cluster is defined as a codeword. All the centers clustered from the training frames produce the codebook for the pLSA model. A frame in the training videos or in the test videos is assigned to a specific codeword in the codebook which has the minimal Euclidean distance to the frame. In the end, a video is encoded in a bag-of-words way, that is, a video is represented using a histogram of codewords, removing the temporal information.

4. pLSA-Based Human Action Recognition

pLSA is a statistical generative model that associates documents and words via the latent topic variables, which represents each documents as a mixture of topics. Our approach uses the bag of words representation as in papers [14–16]. What's difference is that we use the local spatial-temporal maximum value of hierarchical area model, relative velocity and relative acceleration as our features. We suppose that

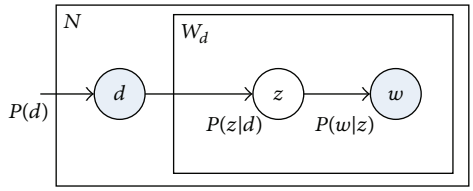


Figure 2: Graph model of pLSA.

the words are independent of the temporal order but related to the spatial order, for the k-means clustering approach with all of the features may lead to the mismatch of the words. Similar local features appearing at different position may be clustered together. When we calculate the frequency of the words, the mismatch appears. And this phenomenon may reduce the precision of the classify approach. In order to solve the problem, we assign spatial information to each word. In the classify approach, we use the pLSA models to learn and recognize human action.

In the context of action categorization, the topic variable z_k correspond to action categories, and each video d_i can be treated as a collection of space-time words w_j. The joint probability of video d_i, action category z_k and space-time word w_j can be expressed as

$$p\left(d_i, z_k, \omega_j\right) = p\left(\omega_j \mid z_k\right) p\left(z_k \mid d_i\right) p\left(d_i\right), \quad (14)$$

where $p(\omega_j \mid z_k)$ is the probability of word ω_j occurring in action category z_k, $p(z_k \mid d_i)$ is the probability of topic z_k occurring in video d_i, and $p(d_i)$ can be considered as the prior probability of d_i. The conditional probability of $p(\omega_j \mid d_i)$ can be obtained by marginalizing over all the topic variables z_k:

$$p\left(\omega_j \mid d_i\right) = \sum_k p\left(z_k \mid d_i\right) p\left(\omega_j \mid z_k\right). \quad (15)$$

Denote $n(d_i, \omega_j)$ as the occurrence of word ω_j in video d_i, the prior probability $p(d_i)$ can be modeled as

$$p\left(d_i\right) \propto \textstyle\sum_j n\left(d_i \mid w_j\right). \quad (16)$$

A maximum likelihood estimation of $p(\omega_j \mid z_k)$ and $p(z_k \mid d_i)$ is obtained by maximizing the function using the Expectation Maximization (EM) algorithm, which the graph model is shown in Figure 2. The objective likelihood function of the EM algorithm is:

$$L = \prod_i \prod_j p\left(\omega_j \mid d_i\right)^{n\left(w_j, d_i\right)}. \quad (17)$$

The EM algorithm consists of two steps: an expectation (E) step computes the posterior probability of the latent variables, and a maximization (M) step maximizes the completed data likelihood computed based on the posterior probabilities obtained from E-step. Both steps of the EM algorithm for pLSA parameter estimate are listed below.

E-step: given $p(w_j \mid z_k)$ and $p(z_k \mid d_i)$ estimate $p(z_k \mid d_i, w_j)$

$$p\left(z_k \mid d_i, w_j\right) \propto p\left(w_j \mid z_k\right)\ p\left(z_k \mid d_i\right). \quad (18)$$

M-step: given the estimated $p(z_k \mid d_i, w_j)$ in E-step, and $n(d_i, w_j)$, estimate $p(w_j \mid z_k)$ and $p(z_k \mid d_i)$

$$\begin{aligned} p\left(w_j \mid z_k\right) &\propto \sum_i n\left(d_i, w_j\right) p\left(z_k \mid d_i, w_j\right), \\ p\left(z_k \mid d_i\right) &\propto \sum_i n\left(d_i, w_j\right) p\left(z_k \mid d_i, w_j\right). \end{aligned} \quad (19)$$

For the task of human motion classification, our goal is to classify a new video to a specific activity class. During the inference stage, given a testing video test, the document specific coefficients $p(z_k \mid d_{\text{test}})$.

We can treat each aspect in the pLSA model as one class of activity. So, the activity categorization is determined by the aspect corresponding to the highest $p(z_k \mid d_{\text{test}})$. The action category k of d_{test} is determined as

$$k = \arg\max_k p\left(z_k \mid d_{\text{test}}\right). \quad (20)$$

In this paper, we treat each frame in a video as a single word and a video as a document. The probability distribution $p(z_k \mid d_{\text{test}})$ can be regarded as the probability of each class label for a new video. The parameter in the training step defines the probability of a word w_j drawing from an aspect z_k. The aforementioned standard EM training procedure for pLSA is to replace

$$p\left(z_k \mid d_i, w_j\right), \qquad p\left(w_j \mid z_k\right), \quad (21)$$

with their optimal possible values at each iteration.

For action recognition with large amount of training data, this would result in long training time. This paper presents an incremental version of EM to speed up the training of PLSA without sacrificing performance accuracy. Assuming the observed data are independent of each other, we propose an incremental EM algorithm presented in Algorithm 1.

Algorithm 1. Incremental EM Algorithm for PLSA Parameter Estimation is as follows.

(1) Inputs;

(2) K—the number of action categories;

(3) D—the number of training videos;

(4) S—the number of videos in each subset;

(5) M—the size of the codebook of spatial-temporal words;

(6) Outputs;

(7) $\widehat{U} = \{\widehat{p}(z_k \mid d_i)\}_{k,i}$;

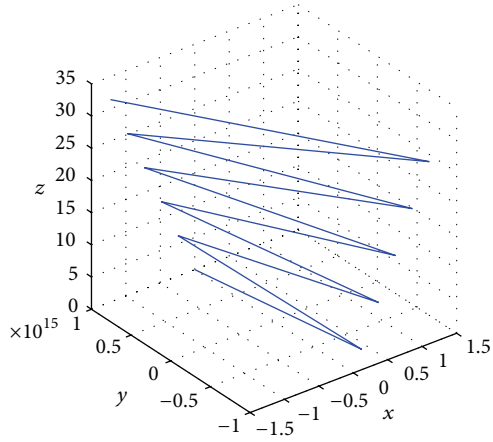
(8) $\widehat{V} = \{\widehat{p}(w_j \mid z_k)\}_{j,k}$;

(9) E-Step;

for all k and j, calculate

$$p\left(w_j \mid z_k\right) = \frac{n_{j,k}}{n_k}. \quad (22)$$

(a)



(b)

	Torso(axis)	Head	Right shoulder	Left shoulder	Right elbow	Left elbow	Right hand	Left hand	Right hip	Left hip	Right knee	Left knee	Right ankle	Left ankle	Right foot	Left foot
Torso(axis)	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Head	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Right shoulder	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Left shoulder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Right elbow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Left elbow	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Right hand	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Left hand	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Right hip	0	0	0	0	0	0	0	0	0	±1	0	0	0	0	0	0
Left hip	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	0	0
Right knee	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	0	0
Left knee	0	0	0	0	0	0	0	0	0	0	−1	0	0	0	0	0
Right ankle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Left ankle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Right foot	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Left foot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(c)

Figure 3: Process of restoring missing coordinate position Remarks: Figure 3 the original picture, Figure 3 Reconstruction of knee motion used the method of least squares data fitting in order to restore missing coordinate position, Figure 3 the occlusion diagram. In the diagram, occlusion part pairs, occlusion state value −1 (red cell for occluded one) and 1 (green cell for occluder), ±1 (orange red cell for rigid body), respectively. In this manner, every part pairs get corresponding occlusion state values.

For all (d_{test}, w_j) pairs and $k \in \{1, \ldots, K\}$ calculate

$$p\left(z_k \mid d_{\text{test}}, w_j\right) = \frac{p\left(w_j \mid z_k\right) p\left(z_k \mid d_{\text{test}}\right)}{\sum_{i=1}^{K} p\left(w_j \mid z_i\right) p\left(z_i \mid d_{\text{test}}\right)}; \quad (23)$$

M-Step: calculate the following:

$$p\left(z_k \mid d_{\text{test}}\right) = \frac{\sum_{j=1}^{N} n\left(d_{\text{test}}, w_j\right) p\left(z_k \mid d_{\text{test}}, w_j\right)}{n\left(d_{\text{test}}\right)}; \quad (24)$$

(10) Repeat E-steps and M-step until the convergence condition is met;

(11) Calculate activity class

$$k = \arg\max_k p\left(z_k \mid d_{\text{test}}\right). \quad (25)$$

5. Experimental Result

5.1. Datasets. We test our algorithm on two datasets: the Weizmann human motion dataset [17], the KTH human action dataset [18, 19], and the HumanEva dataset [3, 20]. All the experiments are conducted on a Pentium 4 machine with 2 GB of RAM, using the implementation on MATLAB. The dataset and the related experimental results are presented in the following sections.

KTH datasets is provided by Schuldt which contains 2391 video sequences with 25 actors showing six actions. Each action is performed in 4 different scenarios.

The WEIZMANN datasets is provided by Blank which contains 93 video sequences showing nine different people, each performing ten actions, such as run, walk, skip, jumping-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop sideways, wave-two-hands, wave-one-hand and bend.

The HumanEva dataset [3, 20] is used for evaluation. It contains six different motions: Walking, Jogging, Gestures, Boxing, and Combo.

In order to evaluate and fairly compare the performance, we use the same experimental setting as in [21, 22]. For every dataset, 12 video sequences taken by four subjects (out of the five) are used for training, and the remaining three videos for testing. The experiments are repeated five times.

The performance of different methods is shown using the average recognition rate. We report the overall accuracy on three datasets. In order to evaluate the performance of occlusion state estimation and reconstruct missing coordinate position, we hand-labeled the ground truth of the occlusion states for test motions. Figure 3 shows how the ground truth of occlusion state is specified.

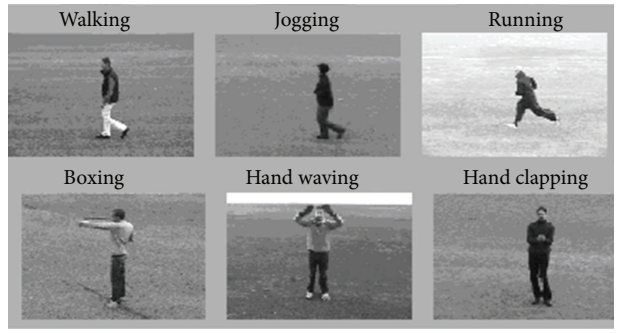
5.2. Comparison. KTH Dataset. It contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoorswith scale variation, outdoors with different clothes, and indoors. Representative frames of this dataset are shown in Figure 4(a). After the process of restoring missing coordinateposition, we use the proposed method, theclassification results of KTH dataset obtained by this approach are shown in Figure 5 and indicate quite a small number of videos are misclassified, particularly, the actions, "running" and "handclapping," are more tended to be confused.

The Weizmann Dataset. The Weizmann human action dataset contains 83 video sequences showing nine different people, andeach performing nine different actions: bending (a1), jumping jack (a2), juming forward on two legs (a3), jumping in place on two legs (a4), running (a5), galloping sideways (a6), walking (a7), waving one hand (a8), waving two hands (a9).

The figures were tracked and stabilized by using the background subtraction masks that come with this data set. Some sample frames are shown in Figure 4(b). The classified results achieved by this approach are shown in Figure 6.

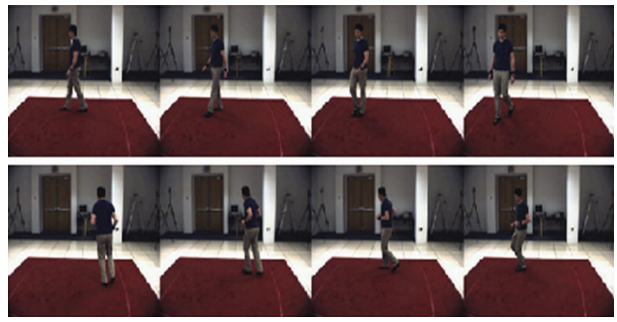
The HumanEva Dataset. The HumanEva dataset is used for evaluation, which are shown in Figure 4(c). It contains five different motions: Walking (a1), Jogging (a2), Gestures (a3), Boxing (a4), and Combo (a5). Each motion is performed by four subjects and recorded by seven cameras (three RGB and four gray scale cameras) with the ground truth data of human joints. The classified results achieved by this approach are shown in Figure 7.

In this paper, we identify jogging, running, walking and boxing and compare the proposed method with the four state-of-the-art methods in the literature: Blank et al. [18], Lu et al. [19], Sigal et al. [3], Chang et al. [20] and Juan Carlos Niebles [21] in three dataset. As shown in the Tables 1, 2 and 3, the existing methods, the low recognition accuracy because these action are not only occlusion situation are complex, but also the legs have complex beat, motion and other group actions. The proposed method can overcome these problems,



(a)

(b)



(c)

Figure 4: Sample frames from our datasets. The action labels in each dataset are as follows (a) KTH data set: walking (a1), jogging (a2), running (a3), boxing (a4), and handclapping (a5); (b) Weizmann data set: running, walking, jumping-jack, waving-two-hands, waving-one-hand, and bending; (c) HumanEva dataset: walking(a1), jogging (a2), gestures (a3), boxing (a4), and combo (a5). Each motion is performed by four subjects and recorded by seven cameras (three RGB and four gray scale cameras) with the ground truth data of human joints.

a1	0.91	0.00	0.03	0.00	0.00
a2	0.00	1.00	0.00	0.00	0.00
a3	0.00	0.00	0.85	0.00	0.00
a4	0.00	0.00	0.00	1.00	0.00
a5	0.03	0.03	0.00	0.01	0.75
	a1	a2	a3	a4	a5

Figure 5: Confusion matrix for KTH data set.

a1	1.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
a2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
a3	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00
a4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
a5	0.03	0.03	0.00	0.01	0.75	0.00	0.31	0.00	0.00
a6	0.00	0.00	0.00	0.00	0.05	0.92	0.04	0.00	0.00
a7	0.00	0.00	0.00	0.00	0.41	0.00	0.95	0.00	0.00
a8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
a9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	a1	a2	a3	a4	a5	a6	a7	a8	a9

Figure 6: Confusion matrix for Weizmann data set.

a1	0.92	0.00	0.03	0.00	0.00
a2	0.00	0.97	0.00	0.00	0.00
a3	0.00	0.00	0.85	0.00	0.00
a4	0.00	0.00	0.00	1.00	0.00
a5	0.03	0.03	0.00	0.01	0.86
	a1	a2	a3	a4	a5

Figure 7: Confusion matrix for HumanEva data set.

Table 1: Compared with other approaches on KTH dataset.

Method	Average recognition rate (%)
The proposed method	92.50
Lu et al. [19] and Blank et al. [18]	81.50
Chang et al. [20] and Sigal et al. [3]	91.20
Niebles et al. [21]	87.04

Table 2: Compared with other approaches on Weizmann dataset.

Method	Average recognition rate (%)
The proposed method	90.10
Lu et al. [19] and Blank et al. [18]	89.30
Chang et al. [20] and Sigal et al. [3]	86.20
Niebles et al. [21]	88.6

and the recognition accuracy and average accuracy are higher than the comparative method.

The experimental results show that the approach proposed in the paper can get satisfactory results and significantly performs better compared the average accuracy with that in [3, 18–21], because of a practical method adopted in the paper.

6. Conclusions and Future Work

In this paper, we proposed an adaptive occlusion state estimation method for 3D human body movement.

Our method successfully recognize without assuming a known and fixed depth order. The proposed method can infer state variables efficiently because it separates the estimation procedure into body configuration estimation and occlusion state estimation. More specifically, in the occlusion state estimation step, at first, we reconstruct human trajectory reconstruction which representing the 3D human pose occlusion relationship and detect body parts having an occlusion relationship using the overlapping body parts by using a Markov random field (MRF) with a state variable. Finally, we use the topic model of pLSA to classify. Experimental results showed that the proposed method successfully estimates the occlusion states in the presence of self-occlusion and the average accuracy is about 92.5%, 90.1%, and 91.4% on the KTH dataset, Weizmann dataset, and HumanEva dataset respectively, which is better than other approaches [3, 18–21].

Table 3: Compared with other approaches on HumanEva dataset.

Method	Average recognition rate (%)
The proposed method	91.40
Lu et al. [19] and Blank et al. [18]	88.70
Chang et al. [20] and Sigal et al. [3]	90.20
Niebles et al. [21]	90.6

We conjecture that the proposed method can be extended for tracking poses from (two or more) interacting people. Tracking poses of interacting people, however, will involve more complex problems such as dealing with more variable motion, inter-person occlusions, and possible appearance similarity of different people.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper (such as financial gain).

Acknowledgments

This research work was supported by the Grants from the Natural Science Foundation of China (no. 50808025) and the Doctoral Fund of China Ministry of Education (Grant no. 20090162110057).

References

[1] X. I. A. Li-min, Q. Wang, and W. U. Lian-shi, "Vision based behavior prediction of ball carrier in basketball matches," *Journal of Central South University of Technology*, vol. 19, no. 8, pp. 2142–2151, 2012.

[2] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[3] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[4] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a pose: tracking people by finding stylized poses," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 271–278, June 2005.

[5] H. Jiang and D. R. Martin, "Global pose estimation using non-tree models," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[6] M. W. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 27–38, 2009.

[7] P. Guo, Z. Miao, Y. Shen, and H.-D. Cheng, "Real time human action recognition in a long video sequence," in *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based (AVSS '10)*, pp. 248–255, Boston, Mass, USA, September 2010.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[9] Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.

[10] B. W. Sy, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1521–1527, June 2006.

[11] N.-G. Cho, A. L. Yuille, and S.-W. Lee, "Adaptive occlusion state estimation for human pose tracking under self-occlusions," *Pattern Recognition*, vol. 46, no. 3, pp. 649–661, 2013.

[12] A. Asthana, M. Delahunty, A. Dhall, and R. Goecke, "Facial performance transfer via deformable models an-d parametric correspondence," *Proceeding of Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1511–1519, 2012.

[13] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, UK, 2nd edition, 2004.

[14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS*, pp. 65–72, October 2005.

[15] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Recognizing human actions by fusing spatio-temporal appearance and motion descriptors," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pp. 3569–3572, Cairo, Egypt, November 2009.

[16] J. Wu and J. M. Rehg, "CENTRIST: a visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[17] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 32–36, August 2004.

[18] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1395–1402, Beijing, China, October 2005.

[19] W.-L. Lu, K. Okuma, and J. J. Little, "Tracking and recognizing actions of multiple hockey players using the boosted particle filter," *Image and Vision Computing*, vol. 27, no. 1-2, pp. 189–205, 2009.

[20] J.-Y. Chang, J.-J. Shyu, and C.-W. Cho, "Fuzzy rule inference based human activity recognition," in *Proceedings of the IEEE International Conference on ControlApplications (CCA '09)*, pp. 211–215, St Petersburg, Russia, July 2009.

[21] J. Carlos Niebles, C. -W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the 11th European Conference Computer Vision (ECCV '10)*, vol. 6312 of *LNCS*, pp. 392–405, 2010.

[22] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Recognizing human actions by fusing spatio-temporal appearance and motion descriptors," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pp. 3569–3572, Cairo , Egypt, November 2009.