Research Article

The Complex Action Recognition via the Correlated Topic Model

Hong-bin Tu,¹ Li-min Xia,¹ and Zheng-wu Wang²

¹ School of Information Science and Engineering, Central South University, ChangSha, Hunan 410075, China

² School of Traffic and Transportation Engineering, ChangSha University of Science & Technology, ChangSha, Hunan 410004, China

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

Received 1 October 2013; Accepted 5 December 2013; Published 16 January 2014

Academic Editors: F. Fernández de Vega and P.-A. Hsiung

Copyright © 2014 Hong-bin Tu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Human complex action recognition is an important research area of the action recognition. Among various obstacles to human complex action recognition, one of the most challenging is to deal with self-occlusion, where one body part occludes another one. This paper presents a new method of human complex action recognition, which is based on optical flow and correlated topic model (CTM). Firstly, the Markov random field was used to represent the occlusion relationship between human body parts in terms of an occlusion state variable. Secondly, the structure from motion (SFM) is used for reconstructing the missing data of point trajectories. Then, we can extract the key frame based on motion feature from optical flow and the ratios of the width and height are extracted by the human silhouette. Finally, we use the topic model of correlated topic model (CTM) to classify action. Experiments were performed on the KTH, Weizmann, and UIUC action dataset to test and evaluate the proposed method. The compared experiment results showed that the proposed method was more effective than compared methods.

1. Introduction

Automatic recognition of human actions from video is a challenging problem that has attracted the attention of researchers in 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. Human complex action recognition is an important research field of the action recognition. Among various obstacles to human complex action recognition, one of the most challenging is to deal with “self-occlusion”, where one body part occludes another. Adaptive self-occlusion behavior recognition has been traditionally tackled by applying statistical prediction and inference methods [2]. 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 and inconsistent track labeling.

Our models are motivated by the recent success of “bag-of-words” representations for object recognition problems in computer vision. The common paradigm of these approaches consists of extracting local features from a collection of images, constructing a codebook of visual words by vector quantization, and building a probabilistic model to represent the collection of visual words. While these models of an object as a collection of local patches are certainly not “correct” ones, (e.g., they only model a few parts of objects and often ignore many structures), they have been demonstrated to be quite effective in object recognition tasks [3–5]. For the recognition approaches, some structures have been lost by moving to this representation. However, this representation is much simpler than the one that explicitly models temporal structures. There has been previous work (e.g., Yamato et al. [6], Bobick and Wilson [7], and Xiang and Gong [8]) that tries to model the full dynamics of videos using sophisticated probabilistic models (e.g., hidden Markov models and dynamic Bayesian networks). Li and Perona [9] use a variant of LDA for natural scene categorization. Sivic et al. [10] perform unsupervised learning of object categories using variants of the pLSA model. In this models, the “words” correspond to local patches extracted by interest point operators, and the “topics” correspond to the different object categories. Fergus et al. [11]

extend pLSA to incorporate spatial information in a translation and scale-invariant manner and apply them to learn object categories from Google's image search. Wang et al. [12] designs a simultaneous classification and annotation framework which extends from LDA and allows image feature and text word share the same dimensional topic space. Putthividhya et al. [13] propose a more general and flexible annotation model which allows different topic spaces for image feature and text word. Bissacco et al. [14] use LDA for human detection and pose classification. The "visual words" in their model are vector quantizations of histogram of oriented gradients in the training images. Niebles et al. [15] recently demonstrate some impressive results on unsupervised learning of human action categories using pLSA and LDA models for human action recognition. Wong et al. [16] adopt pLSA models to capture both semantic (content of parts) and structural (connection between parts) information for recognizing actions and inferring the locations of certain actions.

Optical flow-based action detection methods are well known [17–19]. Efros et al. [20] recognize human actions at a distance in low resolution by introducing a motion descriptor based on optical flow measurements. Ahmad and Lee [21] propose a view independent recognition method by using the C-artesian component of optical flow velocity and human body shape feature vector information. Usually optical flow is used with other features, because it is noisy and inconsistent between frames [22]. Optical flow histograms have also been used to analyze the motion of individual behavior videos. The time series of histogram of optical flow has been modeled as a nonlinear dynamical system using Binet-Cauchy kernels in [23]. However, this approach cannot deal with large motion, for example, a rapid move across frames.

In order to overcome the shortcomings mentioned above, we propose an adaptive self-occlusion action recognition method that not only estimates the occlusion states of body parts but also recognizes the occlusion behavior. Firstly, the Markov random field was used to represent the occlusion state of human body parts. Secondly, the structure from motion (SFM) is used for reconstructing the missing data of point trajectories. Then, we can extract the key frame based on motion feature from optical flow and the ratios of the width and height are extracted by the human silhouette. Finally, we use the topic model of correlated topic model (CTM) for action recognition. Experiments were performed on the KTH, Weizmann, and UIUC action dataset to test and evaluate the proposed method. The experiment results have shown that the proposed method is effective in action recognition.

The reminder of this paper is organized as follows. Section 2 presents the adaptive occlusion state estimation by Markov random field (MRF). In Section 3, we reconstruct the missing data of point trajectories by SFM (structure from motion). Section 4 explains feature representation. Section 5 explains algorithm of action models and the design of the classifier. Section 6 explains the results and analysis of the proposed approach. Finally, we conclude the paper in Section 7.

2. The Adaptive Occlusion State Estimation

The human body is divided into 15 key points, namely, 15 joint points representing the human body structure (torso, pelvis, left upper leg, left lower leg, left foot, rightupper leg, right lower leg, right foot, left upper arm, left lower arm, left hand, rightupper arm, right lower arm, right hand, and head) [24], which represent the human body behavior. In order to calculate the observation, spatial relations, and the motion relationship, we use Markov random field (MRF), which can determine the occlusion positions of the body joints.

In this paper, we use a state variable in the Markov random field (MRF) for representing the self-occlusion relationship between body parts. The MRF is a graph $G = (V, E)$, where V is the set of nodes and E was the set of edges. The graph nodes V represent the state of a human body part and graph edges E model the relationships between the parts [25]. The probability distribution over this graph was specified by the set of potentials defined over the set of edges. The MRF structural parameters [24, 25] are defined as follows:

$X_i = (x_i, y_i, z_i)$: the i th joint point coordinates; $X = \{X_1, X_2, \dots, X_{15}\}$: extract the key points of the body 15, $\gamma(X_i)$ ($i \leq 15$): the i th joints visible parts, and this parameter is used to determine occlusion relation between nodes. When occlusion occurred, trajectories intersected between

$$X_i(X_i(x_i, y_i, z_i)), \quad X_j(X_j(x_j, y_j, z_j)); \quad (1)$$

$\Lambda = \{\Lambda_{ij}\}$ ($i \leq 15, j \leq 15$): the occlusion relation among the 15 body joints. When $\Lambda_{i,j} = 0$, the i th and j th joints are not occluded. When $\Lambda_{i,j} = 1$, the i th occluded j th. When $\Lambda_{i,j} = -1$, the j th occluded i th; $\lambda_i = \{\lambda_1, \dots, \lambda_{15}\}$: the i th occluded joints node.

We apply the MRF model presented in [24, 25] to optimally estimate potential of kinematic relationship and similar to [26] for calculating three potential functions in video activity analysis. The potential of kinematic relationship and three potential functions were defined as follows.

(1) Kinematic relationship is calculated

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

This function indicates the position of two adjacent joints and the angles among joints. $d(x_i, x_j)$ is the Euclidean distance between two adjacent joints. $N(\cdot)$ is the normal distribution with $\mu_k = 0$ and standard deviation δ_k .

(2) The Potential Functions

The observation potential function is

$$\phi_i(I, X_i; \Lambda_i) = \phi_i^C(I, X_i; \Lambda_i) + \phi_i^E(I, X_i; \Lambda_i), \quad (3)$$

where $\phi_i(I, X_i; \Lambda_i)$ is the potential of observation, $\phi_i^C(I, X_i; \Lambda_{ij})$ is the potential of the color, and $\phi_i^E(I, X_i; \Lambda_i)$ is the potential of the edge, I : input image.

The potential of the color is

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

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 as follows:

$\phi_i^{C_{\text{visible}}}(I, X_i; \wedge_{ij})$: the motion state of X_i (the i th body joint) in the viewing area;
 $\phi_i^{C_{\text{occluded}}}(I, X_i; \wedge_{ij})$: the motion state of X_i (the i th body joint) in the occluded area.

The visible term is formulated as

$$\begin{aligned} & \phi_i^{C_{\text{visible}}}(I, X_i; \lambda_i) \\ &= \prod_{u \in (\gamma(X_i) - (\gamma(X_i) \cap \gamma(X_j)))} P_C(I_u) \end{aligned} \quad (5)$$

$$= \prod_{u \in (\gamma(X_i) - (\gamma(X_i) \cap \gamma(X_j)))} \frac{P(I_u | \text{foreground})}{P(I_u | \text{background})}$$

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

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

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

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

$u \in (\gamma(X_i) \cap \gamma(X_j))$: the occlusion area is determined by the calculated overlapping region of X_i and X_j and N is the sum of all occlusion nodes. $f(\theta_i, \theta_j) = 1, T_{\text{lower}} \leq \theta_i - \theta_j \leq T_{\text{upper}}$, 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, the potential of temporal relationship is calculated as follows:

$$\psi_i^T(X_i^t, X_i^{t-1}) = p(X_i^t - X_i^{t-1}; \mu_i, \Sigma_i), \quad (8)$$

where μ_i is the dynamics of X_i at the previous time step and Σ_i is a diagonal matrix with a diagonal element which is identical to $|\mu_i|$, 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 $\wedge^{1:\tau}$ are

$$\begin{aligned} & p(X^\tau | I^{1:\tau}; \wedge^{1:\tau}) \\ &= \frac{1}{Z} \exp \left\{ - \sum_{i \in X^{1:\tau}} \phi_i^C(I, X_i; \lambda_i) - \sum_{ij \in E_k^{1:\tau}} \psi_{ij}^K(X_i, X_j) \right. \\ & \quad \left. - \sum_{i \in E_k^{1:\tau}, t \in 1:\tau} \psi_i^T(X_i^t, X_i^{t-1}) \right\}, \end{aligned} \quad (9)$$

where Z is a normalization constant.

In other words, we put $\phi_i, \psi_{ij}^K, \psi_i^T$ into (4) and get body occluded joints positions:

$$\widehat{X}^t = \arg \max_{X^t} p(X^t | I^{1:t}; \widehat{\wedge}^{t-1}), \quad (10)$$

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

The occluded relation among joints can be obtained by formula (2) as follows:

$$\widehat{\wedge}_{i,j}^t = \arg \max_{\wedge_{ij} \neq 0} \phi_i(I^t, \widehat{X}_i; \wedge_{ij}), \quad (11)$$

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

Therefore, the occluded positions can be calculated by MRF at the entire time of motion.

3. Reconstruct the Missing Data of Point Trajectories

We use the SFM (structure from motion) model to reconstruct the missing trajectories of the occluded joints [27]. Consider a set of P_n point trajectories extracted from the n parts of human body that rigidly moves in f frames. By stacking each image trajectory in a single matrix w_n of $2F \times P_n$, it is possible to express the global motion, which represents the complete trajectory. We define it as

$$w_n = M_n S_n = \begin{bmatrix} R_{1n} t'_{1n} \\ \vdots \\ R_{Fn} t'_{Fn} \end{bmatrix} \begin{bmatrix} X_1 \cdots X_n \\ Y_1 \cdots Y_n \\ Z_1 \cdots Z_n \\ 1 \cdots 1 \end{bmatrix}, \quad (12)$$

where M_n is the $2F \times 4$ human body motion matrix and S_n is the human body contour matrix in homogenous coordinates. Each frame-wise element R_{fn} , for $f = 1, \dots, F$, is a 2×3 orthographic camera matrix that has to satisfy the metric constraints of the model (i.e., $R_{fn} R_{fn}^T = I_2 \times 2$). The 2-vector t'_{Fn} represents the 2D translation of the rigid object (in this paper, we consider the human body as a rigid object). We introduce the registered w'_n measurement matrix such that $w'_n = w_n - t'_{Pn} 1_{Pn}^T$, where 1_{Pn}^T is a vector of P_n ones and $t' = [t'_1, \dots, t'_F]$.

In the case of missing data due to occlusions, we define the binary mask matrix G_n of size $2F \times P_n$ such that 1 represents a known entry and 0 denotes a missing one. In order to solve the components and thus the SFM problem, the equivalent optimization problem [28] can be defined as

$$\begin{aligned} & \text{Minimize} \quad \|G_n \otimes (w_n - M_n S_n)\|^2, \\ & \text{Subject to} \quad R_{fn} R_{fn}^T = I_2 \times 2, \quad f = 1, \dots, F. \end{aligned} \quad (13)$$

Therefore, we can get to make up for the missing point of the trajectory. Figure 1 shows the reconstructed trajectory of the missing point.

4. Feature Representation

4.1. Motion Feature Extraction. The human action can be recognized in terms of hierarchical area model and relative velocity. In this paper, we use optical flow to detect the relative direction and magnitude of environmental motion observed in reference to an observer and also describe the movement of object from current image with respect to the last image. The optical flow [22, 29] equation can be assumed to hold for all pixels within a window centered at p , the local image flow (velocity) vector (V_x, V_y) must be satisfied, and we define some equations as follows:

$$\begin{aligned} I_x(s_1)V_x + I_y(s_1)V_y &= -I_t(s_1), \\ I_x(s_2)V_x + I_y(s_2)V_y &= -I_t(s_2), \\ &\vdots \\ I_x(s_d)V_x + I_y(s_d)V_y &= -I_t(s_d), \end{aligned} \quad (14)$$

where $s_1, s_2, \dots, s_d$ are the pixels inside the window and $I_x(s_i)$, $I_y(s_i)$, and $I_t(s_i)$ are the partial derivatives of the image I with respect to position x , y and time t , evaluated at the point S_i and the current time. These equations can be written in matrix form $Av = b$, where

$$A = \begin{bmatrix} I_x(s_1) & I_y(s_1) \\ I_x(s_2) & I_y(s_2) \\ \vdots & \vdots \\ I_x(s_n) & I_y(s_n) \end{bmatrix}, \quad v = \begin{bmatrix} V_x \\ V_y \end{bmatrix}, \quad (15)$$

$$b = \begin{bmatrix} -I_t(s_1) \\ -I_t(s_2) \\ \vdots \\ -I_t(s_d) \end{bmatrix}.$$

The optical flow vector field V_x is further half-wave rectified into four nonnegative channels.

$F_x^+, F_x^-, F_y^+, F_y^-$, so that

$$F_x = F_x^+ - F_x^-, \quad F_y = F_y^+ - F_y^-. \quad (16)$$

The motion descriptors of two different frames are compared using a version of the normalized correlation.

Suppose the four channels for frame i of sequence A are a_1, a_2, a_3, a_4 and the four channels for frame j of sequence B are b_1, b_2, b_3, b_4 ; then the similarity between frame A_i and B_j is

$$S(A_i, B_j) = \sum_{t \in T} \sum_{d=1}^4 \sum_{x, y \in I} a_c^{i+t}(x, y) b_b^{j+t}(x, y), \quad (17)$$

where T and I are the temporal and spatial extent of the motion descriptors. In this paper, we choose $T = 20$. Therefore, we can obtain the key frames by the frames cluster. Figure 2 depicts results of key-frame extracted by the optical flow.

Then, we extract the feature from the key frame. We assume that the overall key frame number is i , the lengths

of silhouette in horizontal and vertical are $h(i)$ and $v(i)$, respectively, and their ratio is L , so we define it as

$$L = \frac{v(i)}{h(i)}, \quad (18)$$

$$v(i) = m_{\theta+(\pi/2)}(0, 1) \Delta_{\theta+(\pi/2)}.$$

We assume that this area has a maximum value in the θ direction.

The length of the projection area in the $\theta + (\pi/2)$ direction is defined as the area width:

$$h(i) = \frac{\max_{\theta} \{m_{\theta}(0, 1) \Delta_{\theta}\}}{\min_{\theta} \{m_{\theta}(0, 1) \Delta_{\theta}\}}, \quad (19)$$

where Δ is the distance of pixels [31], B is the binary image of the human body area, $m_{\theta}(0, 1)$ is the region B in the perpendicular direction to the parallel lines intersecting the number of lines in group, and Δ_{θ} is the distance between the scan lines. In this paper, we use the eight search directions as shown in Figure 3. Therefore, distances among search lines, respectively, are

$$\begin{aligned} \Delta_0 &= \Delta_{\pi/2} = \Delta, \\ \Delta_{\pi/4} &= \Delta_{3\pi/4} = \frac{\Delta}{\sqrt{2}} \Delta_{\arctg(1/2)} = \Delta_{\arctg 2} \\ &= \Delta_{(\pi/2)+\arctg(1/2)} = \Delta_{(\pi/2)+\arctg 2} = \frac{\Delta}{\sqrt{5}}. \end{aligned} \quad (20)$$

In order to implement the silhouette of human body area, a series of steps must be followed:

- (1) using Gaussian filter to eliminate any noise in video sequence;
- (2) using filter to find the edge strength, which estimates the gradient in the x -direction and the other estimating the gradient in the y -direction;
- (3) the direction of the edge is computed using the gradient in the x and y directions;
- (4) after the edge directions are known nonmaximum suppression now has to be applied. Nonmaximum suppression is used to trace along the edge in the edge direction and suppress any pixel value (sets it equal to 0) that is not considered to be an edge. This will give a thin line in the output;
- (5) threshold $T1$ is applied to the frames, and an edge has an average strength equal to $T1$. Any pixel in the frame that has a value greater than $T1$ is presumed to be an edge and is marked as such immediately and then any pixels that are connected to this edge pixel and that have a value. Figure 4 shows the silhouette of human body area.

Therefore, we can obtain the body contours by this step and use the optical flow descriptor and L descriptor (the ratio of the width and height) to represent the video frames. To construct the codebook, we randomly select a subset from

FIGURE 1: Examples of reconstructed 3-D trajectories. Hip rotations for walking, jogging, and running actions in the KTH dataset [30]. (a) It shows original images. (b) It shows reconstructing the missing data of point trajectories.

FIGURE 2: Results of key-frame extracted method.

FIGURE 3: Eight searching directions in image.

all the frames and compute the affinity matrix on this subset of frames, where each entry in the affinity matrix is the similarity between frame i and frame j calculated using the normalized correlation described above. Then, we run k -medoid clustering on this affinity matrix to obtain V clusters. Code-words are then defined as the centers of the obtained clusters.

4.2. Construct the Codebook. In the end, each video sequence is converted to the “bag of words” representation by replacing

each frame by its corresponding codeword and removing the temporal information.

We follow the instruction from statistical text document analysis, and each image is represented as a bag of code-words. Given a training set of images with annotation words, we use the following notation. Each image is a collection of M visual feature code-words, denoted as $v = \{v_1 : M\}$, where v_m is a unit-basis vector of size V_s with exactly one nonzero entry representing the index of current visual feature in the visual feature dictionary of size V_s . Similarly, for one image annotated with N words $W = \{w_1 : N\}$, we denote each word w_n as a unit-basis vector of size V_t again with only one taking values 1 and 0 otherwise; here V_t is the word dictionary size. Therefore, a collection of D training image word pairs can be denoted as $\{v_{1:D}, W_{1:D}\}$.

5. Action Classification

In order to capture the correlation of topics, we model the hyperparameter of topic prior distribution as multivariate

FIGURE 4: The silhouette of human body area.

FIGURE 5: Graphical representation of the variational distribution for CTM.

normal distribution instead of Dirichlet, similar to [29, 32] and the structure topics of dependencies by covariance matrix, and then we use the logistic normal function:

$$f(\theta_i e) = \left(\frac{\exp \theta_i}{\sum_{j=1}^K \exp \theta_j} \right) \quad (21)$$

to project the multivariate normal to topic proportions for each image; here K is the topic number.

Let $\{\mu, \Sigma\}$ be a K -dimensional mean and variance matrix from normal distribution, and let topics $\pi_{1:K}$ be K multinomials over fixed vocabulary with size V_s and let $\beta_{1:K}$ be K multinomials over a fixed text word vocabulary with size V_t . The CTM generates an image-word pair with M image code-words and N annotation words [29, 32] from the following generative process:

- (1) draw topic proportions $\theta \mid \mu, \Sigma \sim N(\mu, \Sigma)$
- (2) for each visual feature $v_m, m \in \{1, \dots, M\}$
 - (a) draw topic assignment $z_m \sim \text{Mult}(f(\theta))$
 - (b) draw visual feature $v_m \sim \text{Mult}(\pi_{z_m})$
- (3) for each textual word $w_n, n \in \{1, \dots, N\}$
 - (a) draw feature index $y_n \sim \text{Unif}(1, \dots, M)$
 - (b) draw textual word $w_n \sim \text{Mult}(\beta_{z_{y_n}})$.

Firstly, we generate M feature v_m from correlated topic proportions θ , conditional on the topic feature multinomial π , then for each of the N text words, one of the M features is selected and correspondingly assigned to a text word w_n , conditional on the topic word multinomial β . This model is shown as a directed graphical model in Figure 5.

From the generative process of CTM, it could be learned that topic correlations are modeled and generated through the covariance matrix Σ of prior multivariate normal distribution. To learn the parameters of CTM that maximizes

the likelihood of training data, we iteratively estimate the model parameters of latent variables. In this paper, the first step is to add and calculate a set of variational parameters to obtain the approximate lower-bound on likelihood of each sample. The latter step is to estimate the model parameters that maximize the log likelihood of the whole training samples. In the graphical representation of CTM in Figure 5, θ is conditional dependent on the parameter π , which leads to intractability for computing the log likelihood. The mean-field variational distribution is

$$q(\theta, z, y) = q(\theta \mid \gamma, v) \prod_{m=1}^M q(z_m \mid \phi_m) \prod_{n=1}^N q(y_n \mid \lambda_n), \quad (22)$$

where (γ, v) is variational mean and variance of normal distribution, ϕ_m is a variational multinomial over K topics, and λ_n is a variational multinomial over codewords.

Let $\Delta = \{\mu, \Sigma, \pi, \beta\}$ denote model parameters; we bound the log-likelihood of an image-annotation pair (v, w) as follows:

$$\begin{aligned} \log p(v, w \mid \Delta) &= \log \int_{\theta} \frac{p(v, w, \theta, z, y \mid \Delta) q(\theta, z, y)}{q(\theta, z, y)} dx dy \\ &\geq E_q [\log p(v, w, \theta, z, y \mid \Delta)] \\ &\quad - E_q [\log q(v, w, \theta, z, y \mid \Delta)] = \iota(\gamma, v, \phi, \lambda; \Delta), \end{aligned} \quad (23)$$

where E_q is the expectation according to the variational distribution. Taking the log likelihood function $\iota(\gamma, v, \phi, \lambda; \Delta)$ as object function, we fit these parameters with coordinate ascent to maximize the object function:

$$\begin{aligned} \iota &= E_q [\log p(\theta \mid \mu, \Sigma)] + \sum_m E_q [\log p(z_m \mid \theta)] \\ &\quad + E_q [\log p(v_m \mid z_m, \pi)] + \sum_n E_q [\log p(y_n \mid M)] \end{aligned}$$

FIGURE 6: Sample frames from our datasets. The action labels in each dataset are as follows: (a) KTH dataset: walking (a1), jogging (a2), running (a3), boxing (a4), handclapping (a5); (b) Weizmann dataset: running, walking, jumping-jack, waving-two-hands, waving-one-hand, bending; (c) UIUC action dataset: raising two hands (a1), running (a2), clapping (a3), stretching out (a4), and pushing up (a5).

$$+ E_q [\log p(w | y_m, z_m, \beta)] - E_q [\log q(\theta | \gamma, \nu)] - \sum_m E_q [\log q(z_m | \theta)] - \sum_n E_q [\log q(y_m | \lambda)], \quad (24)$$

$$\frac{\partial \iota(\gamma)}{\partial \gamma} = -\sum (\gamma - \mu) + \sum_{m=1}^M \phi_{m,1:K} - \left(\frac{M}{\varsigma} \right) \exp \left(\gamma + \frac{\gamma^2}{2} \right), \quad (25)$$

$$\frac{\partial \iota(\gamma_i)}{\partial \gamma_i^2} = -\frac{\sum_{ii}^{-1}}{2} - \left(\frac{M}{2\varsigma} \right) \exp \left(\gamma + \frac{\gamma_i^2}{2} \right) + \frac{1}{(2\gamma_i^2)}. \quad (26)$$

Then, we update variational multinomial ϕ . The terms in ι with respect to ϕ_m are

$$\iota(\phi_m) = \sum_{i=1}^K \phi_{mi} \left(\gamma_i + \log \pi_{i,v_m} - \log \phi_{mi} + \sum_{n=1}^N \lambda_{nm} \log \beta_{i,w_n} \right), \quad (27)$$

$$\frac{\partial \iota(\phi_{mi})}{\partial \phi_{mi}} = \gamma_i + \log \pi_{i,v_m} + \sum_{n=1}^N \lambda_{nm} \beta_{i,w_n} - \log \phi_{mi} + C, \quad (28)$$

where $\sum_{i=1}^K \phi_{mi} = 1$ is the multinomial constrain. C is the constant.

The full variational inference procedure repeats the updates of (24), (25) until (23); the object function converges.

Then, we will obtain the parameter estimation. Given a collection of human action image data with annotation words $\{v_d, w_d\}_{d=1}^D$, we find the maximum likelihood estimation for parameter $\{\mu, \Sigma, \pi, \beta\}$.

We define $\Delta = \{\mu, \Sigma, \pi, \beta\}$; the overall log likelihood of collection is bounded by

$$\iota = \sum_{d=1}^D \log p(v_d w_d | \Delta) \geq \sum_{d=1}^D \iota_d(\gamma_d, v_d, \phi_d, \lambda_d; \Delta). \quad (29)$$

We maximize the lower boundary of ι , by plugging (23) into (17), and then update model parameters by setting derivation equal zero with respect to each model parameter.

The terms containing π_{ij} are

$$\iota(\pi_{ij}) = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{i=1}^K \sum_{j=1}^{V_s} \phi_{dmi} \log \pi_{i,v_m} - C \left(\sum_{j=1}^{V_s} \pi_{ij} - 1 \right). \quad (30)$$

Setting $\partial \iota(\pi_{ij}) / \partial \pi_{ij} = 0$ leads to

$$\pi'_{ij} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} \phi_{dmi} v_{dm}^j. \quad (31)$$

The terms containing $\beta_{1:K}$ are

$$\iota(\beta_{i,j}) = \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{i=1}^K \sum_{j=1}^{V_t} \sum_{m=1}^{M_d} \phi_{dmi} \lambda_{nm} \log \beta_{i,w_n} - C \left(\sum_{j=1}^{V_t} \beta_{ij} - 1 \right). \quad (32)$$

Setting $\partial \iota(\beta_{ij}) / \partial \beta_{ij} = 0$ leads to

$$\beta'_{ij} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} w_{dn}^j \sum_{m=1}^{M_d} \phi_{dmi} \lambda_{dnm}. \quad (33)$$

The terms containing μ are

$$\iota(\mu) = -\left(\frac{1}{2} \right) \sum_{d=1}^D (\gamma_d - \mu)^T \sum_{d=1}^{-1} (\gamma_d - \mu). \quad (34)$$

a1	0.94	0.03	0.02	0.00	0.00
a2	0.03	0.94	0.07	0.00	0.00
a3	0.02	0.03	0.87	0.00	0.00
a4	0.00	0.00	0.00	0.90	0.25
a5	0.03	0.03	0.08	0.01	0.75
a1	a2	a3	a4	a5	

FIGURE 7: Confusion matrix for KTH dataset.

a1	1.00	0.00	0.03	0.01	0.00	0.00	0.02	0.00	0.00
a2	0.02	0.98	0.01	0.08	0.00	0.00	0.00	0.00	0.00
a3	0.00	0.00	0.75	0.12	0.13	0.00	0.04	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	0.20
a6	0.00	0.00	0.00	0.00	0.05	0.90	0.04	0.00	0.00
a7	0.00	0.15	0.00	0.00	0.41	0.00	0.95	0.00	0.00
a8	0.00	0.00	0.03	0.01	0.00	0.12	0.00	0.98	0.00
a9	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	1.00
a1	a2	a3	a4	a5	a6	a7	a8	a9	

FIGURE 8: Confusion matrix for Weizmann dataset.

Setting $\partial \iota(\mu)/\partial \mu = 0$ leads to

$$\mu' = \left(\frac{1}{D}\right) \sum_{d=1}^D \gamma_d. \quad (35)$$

The terms containing Σ are

$$\begin{aligned} \iota(\Sigma) = \sum_{d=1}^D \left\{ \left(\frac{1}{2}\right) \log |\sigma^{-1}| - \left(\frac{1}{2}\right) \text{Tr} [\text{diag}(\gamma_d^2) \Sigma^{-1}] \right. \\ \left. - \left(\frac{1}{2}\right) (\gamma_d - \mu)^T \Sigma^{-1} (\gamma_d - \mu) \right\}. \end{aligned} \quad (36)$$

Setting $\partial \iota(\Sigma)/\partial \Sigma = 0$ leads to

$$\Sigma' = \left(\frac{1}{D}\right) \sum_{d=1}^D \left(I \times \gamma_d^2 + (\gamma_d - \mu')(\gamma_d - \mu')^T \right). \quad (37)$$

If we obtain variational parameters $\{\gamma_d, \nu_d, \phi_d, \lambda_d, \varsigma_d\}_{d=1}^D$ for all the training samples, we can update the model parameters $\{\mu, \Sigma, \pi, \beta\}$ by plugging them into (30), (34), and (36) until the overall likelihood in (28) converges. Therefore, We approximate the conditional distribution of code-words as follows:

$$p(w | v) = \sum_{m=1}^M \sum_{z_m} q(z_m | \phi_m) p(w | z_m, \beta). \quad (38)$$

The conditional probability $p(w | v)$ can be treated as the predicted confidence score of each annotation word in word vocabulary, given the whole code-words of the unknown human behavior.

6. Experimental Result

6.1. Datasets. We test our algorithm on three datasets: the Weizmann human motion dataset [33], the KTH human

action dataset [30], and the UIUC action dataset [34]. 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.

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

The UIUC action dataset is created by the University of Illinois at Urbana-Champaign (UIUC) in 2008 for human activity recognition. The activities are walking, running, jumping, waving, jumping jacks, clapping, jumping from sit up, raising one hand, stretching out, turning, sitting to standing, crawling, pushing up, and standing to sitting.

For every dataset, 12 video sequences are taken by four subjects (out of the five) 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. In order to evaluate the performance of action recognition, we report the overall accuracy on three datasets.

6.2. 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, outdoors with scale variation, outdoors with different clothes, and indoors. Representative frames of this dataset are shown in Figure 6(a). After process of restoring missing coordinate position, we use the proposed method, and the classification results of KTH dataset obtained by this approach are shown in Figure 7 and indicate quite a small number of videos are misclassified, particularly, the actions “running” and “handclapping” which more tend to be confused.

The Weizmann Dataset. The Weizmann human action dataset contains 83 video sequences showing nine different people and all performing nine different actions: bending (a1), jumping jack (a2), jumping forward on two legs (a3), jumping in place on two legs (a4), running (a5), galloping sideways (a6), walking (a7), waving one hand (a8), and waving two hands (a9). The figures were tracked and stabilized by using the background subtraction masks that come with this dataset. Some sample frames are shown in Figure 6(b). The classified results achieved by this approach are shown in Figure 8.

The UIUC Action Dataset. This dataset consists of 532 high resolution sequences of 14 activities performed by eight actors. The activities are walking, running, jumping, waving, jumping jacks, clapping, jumping from sit-up, raising one

TABLE 1: Compared with other approaches on KTH dataset.

Method	Average recognition rate (%)
The proposed method	90.60
Zhang and Gong [35]	83.40
Gong et al. [36]	89.20
Chang et al. [37]	88.03

TABLE 2: Compared with other approaches on Weizmann dataset.

Method	Average recognition rate (%)
The proposed method	89.20
Zhang and Gong [35]	85.30
Gong et al. [36]	86.30
Chang et al. [37]	82.10

TABLE 3: Compared with other approaches on UIUC dataset.

Method	Average recognition rate (%)
The proposed method	93.30
Zhang and Gong [35]	89.70
Gong et al. [36]	92.50
Chang et al. [37]	91.20

hand, stretching out, turning, sitting to standing, crawling, pushing up, and standing to sitting. Some sample frames are shown in Figure 6(c). The classified results achieved by this approach are shown in Figure 9.

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: Zhang and Gong [35], Gong et al. [36], and Chang et al. [37] in three datasets. As shown in 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, 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 comparing the average accuracy with that in [35–37] because of a practical method adopted in the paper.

7. Conclusions and Future Work

In this paper, we proposed an adaptive occlusion action recognition method for human body movement. Our method successfully recognizes without assuming a known and fixed depth order. We have presented the MRF model and SFM model, which estimates the adaptive occlusion state and recovers the important missing parts of objects in a video clip. This paper presents a new method of human self-occlusion behavior recognition, which is based on optical flow and correlated topic model (CTM). Then, we have employed the optical flow motion feature to extract the key frame and

a1	1.00	0.00	0.03	0.10	0.00
a2	0.00	0.98	0.00	0.15	0.00
a3	0.03	0.04	0.89	0.00	0.10
a4	0.00	0.13	0.00	0.85	0.04
a5	0.04	0.02	0.00	0.01	0.86
	a1	a2	a3	a4	a5

FIGURE 9: Confusion matrix for UIUC dataset.

calculated the ratio of the width and height from human silhouette. Finally, we use the topic model of correlated topic model (CTM) to classify and recognize action. Experiments were performed on the KTH, Weizmann, and UIUC action datasets to test and evaluate the proposed method. The compared experiment results showed that the proposed method was more effective than compared methods and better than other approaches [35–37].

Future work will deal with adding complex event detection to the proposed system, involving more complex problems such as dealing with more variable motion, interperson occlusions, and possible appearance similarity of different people and increasing the database size.

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 work was supported by the National Natural Science Foundation of China (Grant no. 51278068) and by the Science and Technology Project of Hunan (Grant no. 2013GK3012).

References

- [1] L.-M. Xia, Q. Wang, and L.-S. Wu, "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. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.
- [4] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1458–1465, October 2005.
- [5] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 832–838, October 2005.
- [6] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '92)*, 1992.

- [7] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1325–1337, 1997.
- [8] T. Xiang and S. Gong, "Beyond tracking: modelling activity and understanding behaviour," *International Journal of Computer Vision*, vol. 67, no. 1, pp. 21–51, 2006.
- [9] F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 524–531, June 2005.
- [10] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 370–377, October 2005.
- [11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1816–1823, October 2005.
- [12] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1903–1910, June 2009.
- [13] D. Putthividhya, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal Latent Dirichlet Allocation for image annotation," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3408–3415, June 2010.
- [14] A. Bissacco, M.-H. Yang, and S. Soatto, "Detecting humans via their pose," in *Advances in Neural Information Processing Systems*, vol. 19, pp. 169–176, MIT Press, Cambridge, Mass, USA, 2007.
- [15] 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.
- [16] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, June 2007.
- [17] K. Guo, P. Ishwar, and J. Konrad, "Action recognition using sparse representation on covariance manifolds of optical flow," in *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based (AVSS '10)*, pp. 188–195, September 2010.
- [18] U. Mahbub, H. Imtiaz, and M. A. Rahman Ahad, "An optical flow based approach for action recognition," in *Proceedings of the 14th International Conference on Computer and Information Technology (ICCIT '11)*, pp. 646–651, December 2011.
- [19] H. Imtiaz, U. Mahbub, and M. A. R. Ahad, "Action recognition algorithm based on optical flow and RANSAC in frequency domain," in *Proceedings of the 50th Annual Conference on Society of Instrument and Control Engineers (SICE '11)*, pp. 1627–1631, September 2011.
- [20] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the International Conference on Computer Vision*, pp. 726–733, October 2003.
- [21] M. Ahmad and S. Lee, "Human action recognition using multi-view image sequences features," in *Proceedings of the IEEE 7th International Conference on Automatic Face and Gesture Recognition*, pp. 523–528, April 2006.
- [22] U. Mahbub, H. Imtiaz, and A. Rahman Ahad, "Motion clustering-based action recognition technique using optical flow," in *Proceedings of the IEEE/OSA/IAPR International Conference on Informatics, Electronics and Vision*, pp. 919–924, May 2012.
- [23] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1932–1939, June 2009.
- [24] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, UK, 2003.
- [25] 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.
- [26] A. Asthana, M. Delahunty, A. Dhall, and R. Goecke, "Facial performance transfer via deformable models and parametric correspondence," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1511–1519, 2012.
- [27] L. Zappella, A. Del Bue, X. Lladó, and J. Salvi, "Joint estimation of segmentation and structure from motion," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 113–129, 2013.
- [28] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix factorization with missing data," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 316–322, June 2005.
- [29] Y. Wang and G. Mori, "Human action recognition by semilattice topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.
- [30] I. Laptev, *Local spatio-temporal image features for motion interpretation [Ph.D. thesis]*, Computational Vision and Active Perception Laboratory (CVAP), NADA, KTH, Stockholm, Sweden, 2004.
- [31] H. Liu and M. Wang, "Method for classification of apple surface defect based on digital image processing," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 20, no. 6, pp. 138–140, 2004.
- [32] X. Xu, S. Atsushi, and T. Rin ichiro, "Correlated topic model for image annotation," in *Proceedings of The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, pp. 201–208, February 2013.
- [33] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1022–1029, June 2009.
- [34] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proceedings of the 10th European Conference on Computer Vision (ECCV '08)*, pp. 548–561, 2008.
- [35] J. Zhang and S. Gong, "Action categorization by structural probabilistic latent semantic analysis," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 857–864, 2010.
- [36] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771–782, 2011.
- [37] J.-Y. Chang, J. Shyu, and C. Cho, "Fuzzy rule inference based human activity recognition," in *Proceedings of the IEEE International Conference on Control Applications (CCA '09)*, pp. 211–215, Saint Petersburg, Russia, July 2009.