An Algorithm of Unsupervised Posture Clustering and Modeling Based on GMM and EM Estimation

Chuanxu Wang
Faculty of Informatics Qingdao University of Science & Technology, Qingdao, China
E-mail: wangchuanxu_qd@163.com

***Abstract*—this paper focuses on human posture clustering and modeling for human action recognition in the field of computer vision. Specifically we mainly talk about posture description with spatial temporal interesting point features rather than traditional posture segmentations; also we give the comparisons of four kinds of unsupervised clustering methods and continue to carry out unsupervised posture classifications based on Weizmann database. In the following we use GMMs based on EM algorithm to model each clustered posture type. Finally we test our method with Weizmann and KTH Action database. These experiments show its effectiveness and robustness.**

***Index Terms*—unsupervised fuzzy clustering, GMMs, posture modeling, EM estimation**

I. INTRODUCTION

Automatic action recognition is a challenging problem and highlighted in intelligent video surveillance [1-3]. Many methods are proposed on action description and recognition [4], generally there are two types of action description methods, which are low level feature based method and high-leveled human body structure method; while for action recognition there are mainly template matching method and statistical model method. They are summarized respectively as following.

A. Low-Leveled Feature Based Action Description

Low-leveled video feature extraction for action description is somewhat simpler, but it can only represent some limited and inerratic actions. Yilmaz et al. [5] put forward spatial-temporal volume (STV), and analyzed STV by using the differential geometric surface properties to identify action descriptors. Wang et al. [6] proposed silhouette-based Mean Moving Shape (MMS) and moving foreground-based Average Motion Energy (AME) templates for action description, which MMS represents the whole variations of human silhouettes for an action. Additionally, moving trajectory is also extracted as a useful measurement for action modeling; Grimson et al. [7] adopted human moving trajectory and velocity to detect abnormal actions. Robertson et al. [8] regarded that human behavior could be modeled as a stochastic sequence of actions; actions are described by a feature vector comprising both trajectory information (e.g. position and velocity) and a set of local motion descriptors.

B. High-Leveled Feature Based Action Description

The high-leveled human body structure model for action description is mainly associated with human postures in action, which are 2-D posture and 3-D posture. Compared to low-leveled feature methods it is more semantically meaningful and accurate. Ben-Arie et al. [9] adopted 2-D ellipse human structure model, the basic idea is that activities can be positively identified from a sparsely sampled sequence of a few body poses acquired from videos, and an activity was represented by a set of pose and velocity vectors for the major body parts (hands, legs, and torso) which are stored in a set of multidimensional hash tables. 2-D posture is subject to occlusion and view variation problems, so O' Rourke et al. [10] and Gavrila et al. [11] proposed 3-D posture methods. O' Rourke represented a system which was structured as a feedback loop between high and low levels. The domain of human motion lent itself to a model-driven analysis, and the system included a detailed model of the human body. All information extracted from the image was interpreted through a constraint network based on the structure of the human model. Gavrila presented a vision system for the 3-D model based on tracking of unconstrained human movement, using image sequences acquired simultaneously from multiple views; he recovered the 3-D body pose at each time instantly without the use of markers. Roland Kehl et al. [12] presented full body pose tracking using stochastic sampling. A volumetric reconstruction of a person was extracted from silhouettes in multiple video images, then, an articulated body model was fitted to the data with Stochastic Meta Descent (SMD) optimization. All these 3-D posture approaches are expensive in computation for action representation.

C. Template Matching Methods For Action Recognition

The main idea of template matching methods is that the test template is extracted and compared with the reference template at all possible temporal translations. Michael Oren et al. [13] used wavelet template that defined the shape of a person in terms of a subset of the wavelet coefficients of an image, which was invariant to changes in color and texture. Olivier Chomat et al. [14]

doi:10.4304/jsw.6.7.1201-1208

adopted joint statistics of space-time filters to define histograms as templates, which were employed to characterize the activities to be recognized, these histograms provided the joint probability density functions required for recognition using Bayes rule. Shanon X. Ju et al. [15-16] defined a "cardboard person model" in which a person's limbs were represented by a set of connected planar patches. The parameterized image motion of these patches was constrained to enforce articulated motion and was solved for directly using a robust estimation technique. The recovered motion parameters provided a rich and concise description of the activity that can be used for recognition. Bobick and Davis [17] proposed to stack the silhouettes into a Motion Energy Images (MEI) and Motion-History Images (MHI). Seven Hu moments [18] were extracted from both MEI and MHI to serve as action descriptors. Action recognition was based on the Mahalanobis distance between each moment descriptor of the known actions and the input one. Meng [19] extended the MEI and MHI into a hierarchical form and used a Support Vector Machine (SVM) to recognize the actions.

D. Statistical Model Methods For Action Recognition

Statistical models for action recognition are usually learned from some complex training samples, among which Hidden Markov Model (HMM) is typical and well discussed [20-22]. Neil Robertson et al. [1] described actions by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors; action recognition was achieved via probabilistic search of image feature databases representing previously seen actions, and Hidden Markov Models which encoded scene rules were used to smooth sequences of actions. Venkatesh et al. [23] proposed three techniques of feature extraction for person independent action classification in compressed MPEG videos, the feature vectors were fed to Hidden Markov Model for classification of actions， totally seven actions were trained with distinct HMM for classification, and the recognition results of more than 90% have been achieved. Ahmad, M. et al. [2] proposed human action recognition from image sequences in different viewing angles that used the Cartesian component of optical flow velocity and human body shape feature vector information, then performed PCA to reduce the higher dimensional shape feature space, and each action for any viewing direction was modeled using a set of multidimensional discrete hidden Markov model. Bregler et al. [3] proposed that recognition was the succession of very general low level grouping mechanisms to increased specific and learned model based grouping techniques at higher levels. Low-leveled primitives were areas of coherent motion found by EM clustering, mid-leveled categories were simple movements represented by dynamical systems, and high-level complex gestures were represented by Hidden Markov Models as successive phases of simple movements.

E. Orgnizition of This Paper

Study of kinematics for human motion shows that an action can be effectively described with 3 to 5 postures in a proper temporal sequence [24-25]; therefore posture modeling is critical for both action representation and recognition. The above mentioned posture extraction methods mainly rely on human body segmentation or background removing, which are intractable for noise perturbation caused by illumination variation and occlusion and camouflage or shadow. Spatial temporal interesting points (STIPs) [26-27] are regarded effective methods to extract low level motion features without background modeling in the recent years. We propose an unsupervised algorithm to classify salient postures from Weizmann database with STIPs, which is needless segmentation of human silhouettes; further more, GMM based on EM method is adopted to model the clustered posture types.

The rests of the paper are organized as following. Section II details how STIPs are extracted from training action videos and the pose descriptor is put forward. In section III we talk about posture clustering, four unsupervised classification methods are compared, and gives the salient posture clustering results. In section IV we details how each salient posture is modeled with GMM based on EM. Experiments are conducted in both Weizmann and KTH action databases in section V, which is to prove the effectiveness and robustness of our method. Finally we give conclusions and future work in section VI.

II. LOW LEVEL FEATURE EXTRACTION FOR POSTURE

A. Shortcomings of Pose Silhouette Segmentaions

Traditionally segmentation methods are adopted to get the low level features of postures. Generally they are subject to noise interferences from background, such as illuminant variation, dynamic background, camouflage etc., which degrade the semantic meaning of the segmented pose results. Fig.1 shows segmented human object images, which indicate serious interruption from shadows. These pose silhouettes are deformed and lost their geometric meanings; therefore they was degraded as low SNR inputs for next step of posture recognition.

Figure 1. Silhouettes are deformed due to shadows in human object segmentation

B. Extraction of STIPs

Human action features are located at the spatial temporal neighborhood, where the image values have large variations in both the spatial and the temporal dimensions [26]. We can use fewer feature points to identify human's movement behavior, without the need of segmenting and tracking the whole human body any more. Points with such properties will be spatial-temporal points with a

distinct location in time corresponding to the moments with non-constant motion of the image in a local spatial-temporal neighborhood. For example, during the walking process, these interesting neighborhoods locate at feet lifting and landing, knees bending and so on.

There are two methods to extract these spatial-temporal interesting points, which are proposed by Ivan Laptev and Dollár. The method of Ivan Laptev is to detect 3-D Harris corners as STIPs. The extracted STIPs based on Ivan Laptev [26] are shown in Fig.2. They are sparse and sensitive to scales, and not adaptive to posture modeling. So we choose the STIPs extraction method based on Dollár in this paper.

Compared to Ivan Laptev, Dollár's [27] method considered that any region with spatially distinguishing characteristics undergoing a complex motion can induce a

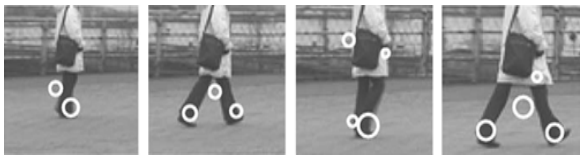


Figure.2 STIP extraction based on Ivan Laptev

strong response. The response function can be calculated as:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (1)$$

Where I is input gray video, and $g(x, y; \sigma)$ is the 2D

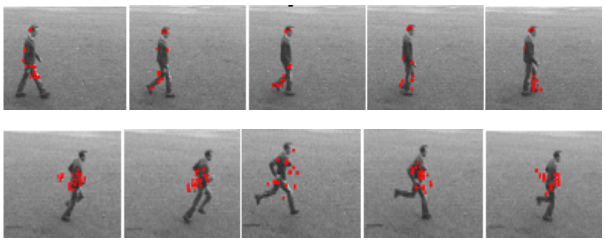


Figure.3 STIPs extraction based on Dollár's method

Gaussian smoothing kernel, applied only along the spatial dimensions, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally. Fig.3 shows STIPs extraction method based on Dollár. We can see that STIPs are located at the region of body's intense movement, so human behavior characteristics can be described by STIPs.

C. Descriptor of STIPs

In order to describe the distributions of STIPs for a posture, we need to design the descriptor of each STIP. Every STIP stands for a small area that is undergoing non-constant movement, so we choose an $5 \times 5 \times 5$ adjacent neighborhood which is called cuboid to model a STIP, where we calculate the 3D gradients (L_x, L_y, L_t) for each pixel, so each STIP can form a 375-dimensioned vector as its descriptor. Here, it is emphasized that 3D gradients (L_x, L_y, L_t) should be flattened as $(L_x/Norm, L_y/Norm, L_t/Norm)$, where $Norm$ is calculated as:

$$Norm = \sqrt{L_x^2 + L_y^2 + L_t^2} \quad (2)$$

D. descriptor for a posture

Because a posture is corresponding to a set of STIPs, then we can describe this pose by its statistic distribution of STIPs, and also we can classify these postures via clustering their distributions of STIPs. In this paper, we model a single posture by calculating the histogram of all its STIPs in a pose. That is, a STIP descriptor is 375-dimension, which is composed 3 gradient sub-vectors, they are 2 spatial gradients on x and y directions and 1 temporal gradient on t direction respectively, each of them is 125-dimension. Suppose there are N STIPs in a posture frame, we calculate a histogram of 16 bins respectively for 3 types of gradient sub-vectors, and then we combine these 3 histograms as a 48-bined histogram as the descriptor of a single pose in a frame.

III. Unsupervised Posture Clustering

A. Unsupervised Fuzzy Clustering Algorithms

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity. Posture similarities is fuzzy, in the framework of fuzzy clustering, it allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters [28]. In fuzzy relational clustering, the problem of classifying data is solved by expressing a relation that quantifies the similarity, or dissimilarity degree between pairs of objects. Based on such relation, objects very similar to each other, i.e., objects of the same type will belong with high membership values to the same cluster [29]. Typically there are fuzzy C-mean methods and some of its modified versions, which can serve as unsupervised posture classification tools to our problems. Before posture clustering experiments it is necessary to compare these related versions of fuzzy C-mean algorithms and evaluate their efficiency, so that we can make an optimal choice for our problems.

B. Pros and Cons of Fuzzy C-means

The popularity and usefulness of fuzzy C-means result from three facts. The algorithms are simple; they are very effective at efficiently finding minimums of objective function $\boldsymbol{J_m}$: give data set

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n\} \quad (3)$$

Where n is the number of data points in $\boldsymbol{X}$, $\boldsymbol{x_k} \in \boldsymbol{R^p}$; $\boldsymbol{p}$ is the number of features in each vector $\boldsymbol{x_k}$; in order to cluster $\boldsymbol{X}$ into C prototypes, $\boldsymbol{J_m}$ is sought as

$$\min_{(U,V)}\left\{J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}D_{ik}^{2}\right\} \quad (4)$$

$$\sum_{i=1}^{c}u_{ik}=1,\forall k$$

Constrain .

And distance $D_{ik}^2=\|x_k-v_i\|_A^2$.

A-norm $\|x_k\|_A=\sqrt{\langle x,x\rangle_A}=\sqrt{x^TAx}$. Degree of fuzzy $m\geq 1$; $v=(v_1,\ v_2,\ \cdots\ v_C)^T$. And U is the membership functions, those minimizers usually represent the structure of X very well; test result is shown in Fig.4. Various theoretical properties of the algorithms are well understood, and are described in Refs [30]. Additionally, this method is unsupervised and always convergent.

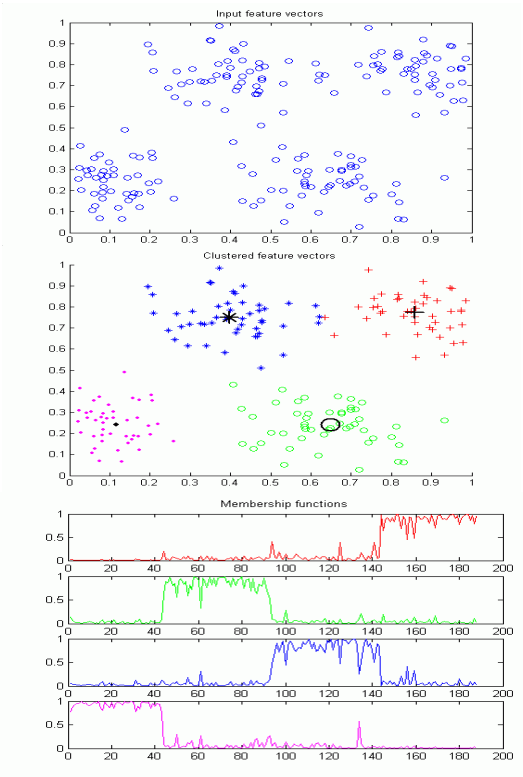


Figure.4 FCM clustering test.
From up to bottom they are input data, clustering result and membership of functions U, Where n=188, p=2, C=4.

Also this method does have some disadvantages, such as, long computational time, sensitive to the initial guess (speed, local minima), unable to handle noisy data and outliers, very large or very small values could skew the mean, not suitable to discover clusters with non-convex shapes.

C. Relational Fuzzy C-means

The relational fuzzy C-means (RFCM) classifier is useful when a feature space has an extremely high dimensionality that exceeds the number of objects and many of the feature values are missing, or when only relational data are available instead of the object data. The relational data is represented by a matrix in terms of distances (dissimilarity) between object data, and is not concerned with the relational database. Of course the pair wise relational matrix can be easily figured out while the data are given as data vector sets. So RFCM can deal with more than the problems that FCM can do.

Whenever relational data are available that corresponds to measures of pair wise distances (actually, squared distances) between objects, RFCM can be used instead which rely on its computation efficiency. One of the advantages is that their driving criterion is "global", i.e. it assesses a property implicitly shared by all the objects even though the object data is not directly used. Another advantage is that these relational algorithms automatically inherit excellent numerical convergence properties of FCM, because they have a close relationship with the quickly convergent and reliable object-oriented algorithms.

Give matrix $R=[r_{ij}]$ for relational data, which is corresponding to pair wise distance between objects, different to FCM, its object function is defined as $K_m(U)$,

$$K_m(U)=\sum_{i=1}^{c}\left(\sum_{j=1}^{n}\sum_{k=1}^{n}(u_{ij}^{m}u_{ik}^{m}\delta_{jk}^{2})/(2\sum_{t=1}^{n}u_{it}^{m})\right) \quad (5)$$

where $m\geq 1$ and for $1\leq j,k\leq n,\delta_{jk}^2=r_{jk}$. Useful partitioning U of the data are sought as minimizers of $K_m(U)$. The optimal partitioning U^* gives

$$K_m(U^*)=J_m(U^*,F_m(U^*))=\min_v J_m(U,v) \quad (6)$$

from which it follows that U* is a minimizer of $K_m(U)$, if and only if U^* is a minimizer of $\min_v J_m(U,v)$, which is easily shown to be true if and only if (U^*,v^*) is a minimizer of $J_m(U,v)$. The above explanation proves that FRCM has preserved the simplicity and convergence of FCM. Additionally, numerical experiments show the actual work done per iteration could be smaller for the RFCM algorithms than for the object data versions FCM, when the dimension p of the feature data is large.

RFCM has a strong restriction which restrains its applications. The relation matrix R must be Euclidean, i.e., there exists a set of N object data points in some p-space whose squared Euclidean distances match values in R. To ease the restrictions that RFCM imposes on the dissimilarity matrix, there are two improved versions of RFCM which are introduced in the following.

D. None Euclidean Relational Fuzzy C-means

None Euclidean Relational Fuzzy (NERF) C-means can transform a non-Euclidean relational matrix into Euclidean ones by using the β -spread transformation introduced in [31]. This transformation consists of adding a positive number β to all off-diagonal elements of R. As proved in [31], there exists a positive number β_0 such that the β -spread transformed matrix R_β is Euclidean

for all $\beta \geq \beta_0$, and is not Euclidean for all $\beta \leq \beta_0$. The parameter β, which determines the amount of spreading, should be chosen as small as possible to avoid unnecessary spreads of data with consequent loss of cluster information.

On the other hand, the exact computation of β_0 involves an expensive eigen value computation [31]. To reduce loss of information without decreasing performance dramatically, Hathaway and Bezdek [31] proposed an extension of RFCM, denoted non-Euclidean RFCM (NERFCM), in which the β -spread transformation is computed dynamically during the iteration process of RFCM. The β_N computed by NERFCM is the minimum value which guarantees the convergence of RFCM. As RFCM can converge even if a relation is not Euclidean, it may happen that $\beta_N < \beta_0$. NERFCM has proved to be one of the most reliable fuzzy relational clustering algorithms; the performance of NERFCM depends, however, on the value of β_N which could be so large that the structure inherent in R might not be mirrored by that in $\boldsymbol{R}_{\beta_N}$ [31].

E. Any Relational Fuzzy C-means

Any Relational Fuzzy C-means (ARCM) represents a cluster in terms of a representative of the mutual relationships of the objects which belong to the cluster with a high membership value. Each object is represented by the vector of its relation strengths with the other objects in the data set, and a prototype is an object whose relationship with all the objects in the data set is representative of the mutual relationships of a group of similar objects. Like FCM, ARCM partitions the data set by minimizing the Euclidean distance between each object (strongly) belonging to a cluster and the prototype of the cluster. ARCM determines the optimal partition by minimizing the following objective function:

$$\boldsymbol{J_m(U,v) = \sum_{i=1}^{C}\sum_{k=1}^{n} u_{ik}^{m}\delta^2(x_k, v_i)} \qquad (7)$$

where $\boldsymbol{\delta(x_k, v_i)}$ is the deviation between, respectively, the relation between x_k and all the other objects, and between v_i and all the other objects.

Defining

$$\boldsymbol{\delta(x_k, v_i) = \sqrt{\sum_{s=1}^{n}(r_{ks} - v_{is})^2}} \qquad (8)$$

Where $\boldsymbol{r_{ks}}$ is the relation between the pair of objects x_k and x_s, and $\boldsymbol{v_{is}}$ is the relation between the prototype v_i and object x_s, and applying the standard Lagrange multipliers minimization method, ARFCM algorithm can get the final convergence and give the clustering membership matrix U.

F. Implimentation of Posture Clustering With NERFCM

It is found that NERFCM is of good property for posture similarity propagation in clustering compared to other 3 methods. Then we carry out posture classification based on this algorithm. We define the pair-wised similarity of two postures with histogram intersection method, which is:

$$S(p,q) = \sum_{u=1}^{B} \min\{p^{(u)}, q^{(u)}\} \qquad (9)$$

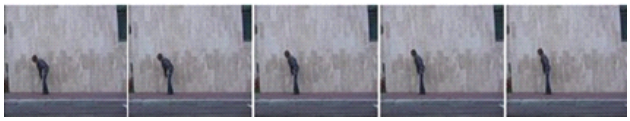
Where p and q are histogram descriptors of 2 postures, and each histogram is of B bins. If they are the same, the similarity s is 1, so the dissimilarity can be defined as $d = 1 - s$. Consequently, the posture dissimilarity of the total N poses in a database can be calculated, and they can form a dissimilarity matrix:

$$D = [d_{ij}]_N = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \qquad (10)$$

Where the value of diagonal elements d_{ii} is 0, the other elements value d_{ij} is the dissimilarity between posture i and j.

The N frames are then clustered into M clusters by employing unsupervised Non-Euclidean Relational Fuzzy (NERF) C-Means. Its classification results are the probabilities of each sample belonging to the M clusters, and by the maximum probability of membership of each for a posture, we can decide their belongings. In this paper we use Weizmann database for learning, there are 4126 poses are used and cluster number M is 37. Some clustering results are show in Fig.5 and Fig.6.

Figure.5 example poses of cluster 8, which are from action Skip and Run, representing for their visual similarities.



(a) Example frames of bending up

(b) Example frames of waving up

Figure.6 example poses from cluster 13, shared by action Wave and Bend consisting of similar movement courses

It intuitively satisfies human senses that pose frames in Fig.5 are classified into the same cluster, because they are indeed of similar postures. But it is puzzled that sample frames in Fig.6 belong to a same cluster, for they are not visually similar. The reason can be explained as following, though bending up and handing up are not

spatially similar, but they are of similar moving up course, that is, these two actions are temporally similar, just because that the STIPs are of temporal property.

IV. Posture Modeling Based on GMM and EM

A. Gaussian Mixture Model

In this paper, Gaussian Mixture Models are regarded as using several distributions to describe each type of clustered postures. In other words, we use K weighted sum of Gaussian distribution functions to close in the distribution function of each posture's observed values.

For a single sample x_i in each type of clustered posture data set $X=\{x_1,x_2,\ldots x_N\}$, the Gaussian mixture distribution density function is:

$$P(x_i|\Theta)=\sum_{i=1}^{K}\omega_k p_k(x_i|\theta_k) \quad (11)$$

Where K is the number of Gaussian distributions, ω_k is the weight estimation of k^{th} Gaussian in the mixture, and it satisfied with: $\sum_{k=1}^{K}\omega_k=1$. p_k is the Gaussian probability density function, and $\Theta=(\theta_1,\theta_2,\ldots,\theta_K)$ is the parameter vector of mixture composition. $\theta_k=(\mu_k,\sigma_k)$ is the Gaussian distribution parameter, that is, the mean value and covariance matrix respectively.

There are two main methods to estimate the parameters of GMM, which are based on the online updating and EM algorithm. The principle of the online updating method can be described as follows. Every new observation x_i is checked with each of K current Gaussian distributions, if it is matched with j^{th} Gaussian model, the parameter $\mu_{j,t}$ and $\sigma^2_{j,t}$ for the matched distribution is updated respectively as:

$$\begin{cases}\mu_{j,t}=(1-\alpha)\cdot\mu_{j,t\text{-}1}+\alpha\cdot I_t\\ \sigma^2_{j,t}=(1-\alpha)\cdot\sigma^2_{j,t\text{-}1}+\alpha\cdot(\mu_{j,t}-I_t)^2\end{cases} \quad (11)$$

$$\omega_{j,t}=(1-\alpha)\cdot\omega_{j,t\text{-}1}+\alpha\cdot M_{j,t} \quad (12)$$

Where α is the Gaussian adaptation learning constant, $M_{n,t}$ is 1 for the j^{th} model which matches new observation x_i, and 0 for the rest none matched models.

The other method of GMM parameter estimation is based on EM. Our sample data is incomplete, and EM algorithm is capable of parameter estimation with MLE (Maximum likelihood Estimation) under insufficient samples. So we choose EM to estimate the parameters of GMM.

D. EM Algorithm

It is an iterative algorithm to get the maximum likelihood estimation of distribution density function, when the observation data is incomplete. It can significantly reduce computational complexity, but the performance is similar with the maximum likelihood estimation, so it has a good practical application value. In this paper, the observation data X of each posture cluster is incomplete, so the missing data Y is introduced, and the complete data is formed as $Z=\{X,Y\}$, where y_i is the class that x_i belongs to. If x_i comes from the k^{th} Gaussian component, we can obtain $y_i=k$. So the likelihood function of complete data is:

$$L(\Theta|Z)=L(\Theta|X,Y)=P(X,Y|\Theta).$$

There are two steps in EM algorithm: Expectation step and Maximization step. When the observation data and the current parameter are known, we can get the Expectation Maximization of complete likelihood function $L(\Theta|Z)$ according to the missing data Y. So the E-step and M-step are:

$$\text{E-step: } Q(\Theta,\Theta^t)=E\left[\log P(X,Y|\Theta)\middle|X,\Theta^t\right] \quad (13)$$

$$\text{M-step: } \Theta^{t+1}=\arg\max Q(\Theta,\Theta^t) \quad (14)$$

Formulas (13) and (14) can ensure to get the maximum after the iterative computation E-step and M-step.

C. EM Estimation on GMM

To complete the algorithm mentioned above, the key step is to get the probability density of the missing data Y. We can get the probability density of Y according to Bayes $p(Y|X,\Theta^t)=\prod_{i=1}^{N}p(y_i|x_i,\Theta^t)$. So the iterative functions based on EM estimation are:

$$\left.\begin{aligned}&\alpha_k^t=\sum_{i=1}^{N}p(k|x_i,\Theta^t)\\&\pi_k^t=\frac{1}{N}\alpha_k^t\\&\mu_k^{t+1}=\frac{1}{\alpha_k^t}\sum_{i=1}^{N}x_i p(k|x_i,\Theta^t)\\&\sum_k^{t+1}=\frac{1}{\alpha_k^t}\sum_{i=1}^{N}x_i p(k|x_i,\Theta^t)(x_i-\mu_v^{t+1})(x_i-\mu_v^{t+1})^T\end{aligned}\right\} \quad (15)$$

In this paper we model each clustered salient posture type with GMM whose parameters e.g. mean μ_k and covariance σ_k, also the weights ω_k, are estimated by EM. The following data is the GMM information for cluster 5, where there are 95 frames in cluster 5, and it is modeled with 3 components of Gaussians, which are:

Nin: 48
Ncentres: 3
Cover_type:"spherical"
Priors: [0.3765 0.2616 0.3618]
Centres: [3x48 double]
Covars: [3.9156e-004 7.2282e-004 5.0298e-004]

The first item "Nin: 48" stands for its dimension is 48; the second item"ncentres: 3" stands for there are 3 components of Gaussians; the third stands for its covariance shape is 'spherical'; the fourth one is its weights, and so on.

V. POSTURE RECOGNITION AND EXPERIMENT RESULTS

A. Posture Recognition Algorithm

In order to test the efficiency of GMMs on a salient posture type, we design the posture recognition experiments base on Weizmann and KTH Action data base. The key question is how to measure the similarity between the GMMS and the test pose sequence. Give a pose frame, its STIPs can be extracted, then we calculate its gradient histogram descriptor f based on the cuboids of each STIP. Posture recognition is to find the best matching for f among all the cluster GMM models. That is to calculate the probability:

$$p(f) = \arg\max_{\varphi \in \psi} \sum_{i=1}^{K} \omega_{i,\varphi} \cdot \eta_{i,\varphi}(f \cdot \mu_i \cdot \sigma_i) \quad (16)$$

Where ψ is the collection of all cluster models, φ is one of ψ, and φ has K sub-Gaussian models of GMM, $\eta_{i,\varphi}$ is the i^{th} Gaussian probability density function, μ_i, σ_i are its mean and variance respectively, $\omega_{i,\varphi}$ is the weight of i^{th} Gaussian model in cluster φ.

If maximum probability $p(f)$ is larger than a threshold, then the input pose frame can be judged belonging to one specific cluster.

B. Experiment Results

We take Weizmann database as training samples, there are 10 types of action and each action has 9 action videos conducted by 9 different persons, in all there are 4126 poses with detected STIPs in our training experiments. After clustering, 4126 frames are clustered into 37 salient posture types.

In order to verify the overall performance of the proposed GMM models, we take Leave One Sample Out Test (LOSO), that is, taking out one action video as test samples from these 10 actions, and the rests are training samples. We tested all these 37 GMMs. In Tab.I we list the test results.

TABLE I.
SALIENT POSTURE MODEL TEST

Salient Posture Types (GMMs)	Test Pose Frames	Correct Recognition Frames	Correct Percentage
1	127	110	0.86614
2	107	100	0.93458
3	60	56	0.93333
4	63	59	0.93651
5	95	89	0.93684
6	46	42	0.91304
7	60	55	0.91667
8	221	190	0.85973
9	85	81	0.95294
10	90	81	0.90000
11	93	88	0.94624
12	33	30	0.90909
13	99	93	0.93939
14	92	85	0.92391
15	147	134	0.91156
16	171	155	0.90643
17	60	54	0.90000
18	37	33	0.89189
19	30	29	0.96667
20	39	39	1.00000
21	43	42	0.97674
22	161	139	0.86335
23	76	72	0.94737
24	35	34	0.97143
25	59	54	0.91525
26	67	58	0.86567
27	145	127	0.87586
28	114	97	0.85088
29	169	164	0.97041
30	54	45	0.83333
31	30	27	0.90000
32	55	52	0.94545
33	41	39	0.95122
34	82	76	0.92683
35	82	79	0.96341
36	233	213	0.91416
37	47	44	0.93617

VI. CONCLUSIONS

Human behavior recognition has been a hot and important topic recently. In this paper, we have proposed an effective algorithm on posture modeling base on GMMs and EM, and it can obtain a high recognition rate. The experiments prove that our method is accurate and effective, which is robust to the interferences caused by video segmentation, such as, illumination variation and camouflage, and so on.

However, there are still some disadvantages. For example, it is only effective for stable camera environment and simple background. Our next step work is to improve our algorithm to adapt to the dynamic camera environment and complex background. In addition, we will make action recognition with Markov models.

ACKNOWLEDGMENT

The authors would like to thank M. Blank at the Weizmann Institute and I. Laptev at Computational Vision and Active Perception Laboratory (CVAP), NADA, KTH, Stockholm for sharing their datasets. This research is also partially supported by Natural Science Fund of Shandong (ZR2010FL007 and ZR2009GM007 and Y2008G09) and the Project of Shandong Province Higher Educational Science and Technology Program (J10LG23 and J09LG12).

REFERENCES

[1] Robertson, N., Reid, I.D, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, vol.104, pp.232-248, 2006.

[2] Ahmad M., Seong-Whan Lee. "Human action recognition using multi-view image sequences features," in *Proc. AF GR*, pp.523-528, 2006.

[3] Christoph Bregler. "Learning and Recognizing Human Dynamics in Video Sequences," *IEEE CVPR*, pp.568-574, 1997.
[4] Junxia Gu, Xiaoqing Ding, Shengjin Wang. "Survey on action analysis," *Journal of china image and graphics,* Vol.13, pp.377-387, 2010.
[5] Yilmaz Alper, Shah Mubarak, "Actions sketch: A novel action representation," In *Proc. CVPR*, San Diego, California, USA, pp.984-989, 2005.
[6] Wang Liang, Suter David, "Informative shape representations for human action recognition," In *Proc. of Conf. on PR*, Hong Kong, pp.1266-1269, 2006.
[7] Grimson W E L, Stauffer C, Romano R, et al, "Using adaptive tacking to classify and monitor activities in a site," *In Proc. CVPR*, Santa Barbara, California, USA, pp.22-29,1998.
[8] Robertson N, Reid I, "Behavior understanding in videos: a combined method," In: *Proc. ICCV*, Beijing, 2005, pp. 808-815.
[9] Ben-Arie Jezekiel, Wang Zhi-qian, Pandit Puvin, et a l, " Human activity recognition using multidimensional indexing," *IEEE Trans. PAMI*, vol.24, no.8, pp.1091-1104, 2002.
[10] Rourke J O, BadlerN, "Model-based image analysis of human motion using constraint propagation," *IEEE Trans. PAMI*, vol.2, no.6, pp.522-536, 1980.
[11] Gavrila Dariu M, Vision-Based 3-D Tracking of Human in Action". Maryland, USA: University of Maryland, 1996.
[12] Kehl R, BrayM, VanGool L. "Full body tracking from multip le views using stochastic sampling," In *Proc. CVPR*, San Diego, California, USA, 2005, pp.129-136.
[13] M.Oren, C.Papageorigiou, P.Sinha,E.Osuna, T.Poggio, "Pedestrian detection using wavelet templates," In *Proc. CVPR*,1997,pp.193-199.
[14] O.Chomat, J.L.Crowley, "Recognizing motion using local appearance," In *proc. on IRS*, University of Edinburgh, 1998.
[15] S.X.Ju, M.J.Black, Y.Yacoob. "Cardboard people: A Parameterized model of articulated image motion," in *Proc. On AFGR*, 1996, pp.38-44.
[16] Y.Yacoob, M.J.Black, "Parameterized modeling and recognition of activities," In: *Proc. ICCV*, 1998, pp.120-127.
[17] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. PAMI*, vol. 23, pp. 257–267, 2001.
[18] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Information Theory*, vol. 8, no. 2, pp. 179-187, 1962.
[19] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *Proc. CVPR*, 2007.
[20] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition," in *Proc. of IEEE*,1989,vol.77, no.2, pp.257-286.
[21] Michael M, Yin, Jason T.L. Wang, "Application of Hidden Markov Models to Gene Prediction in DNA," In *Proc. on I IS*,1999, pp.40-48.
[22] Shun-Zheng Yu, Hisashi Koayashi, "An efficient Forward-Backward Algorithm for an Explicit Duration Hidden Markov Model," *IEEE Signal Processing Letters*, 2003, vol.10, no.1, pp.11-14.
[23] Venkatesh Babu R., Anantharaman B., Ramakrishnan K.R., Srinivasan S.H., "Compressed domain action classification using HMM," in *IEEE workshop on Content-Based Access of Image and Video Libraries*.2001,pp.44-49.
[24] P. Kakumanu, S. Makrogiannis, N. Bourbakis, "A survey of skin-color modeling and detection methods". *Pattern Recognition* vol.40, pp.1106-1122, 2007.
[25] Cheng duansheng, Liukaisheng, "Summarization of skin detection techniques," *Journal of computer*, vol.29 no.2 Feb，2006, pp.194-207.
[26] Ivan Laptev and Tony Lindeberg, "Space-Time Interest Points," In *Proc. ICCV*, 2003, Nice, France, pp.I:432-439.
[27] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatial-temporal features," in *Proc. IEEE International Workshop on VEPRTS*, pp.65-72, 2005.
[28] Bezdek JC, Hathaway RJ, Sabin MJ, Tucker WT , "Convergence theory for fuzzy C-means: counterexamples and repairs," *IEEE Trans. on Systems, Man, and Cybernetics*. Vol. SMC-17, no. 5, pp. 873-877. 1987.
[29] Mingzhou (Joe) Song and Lin Zhang, "Comparison of Cluster Representations from Partial Second- to Full Fourth-Order Cross Moments for Data Stream Clustering," in *Proc. 8th Conference on Data Mining*, pp.560-570. 2008.
[30] Setnes M, Babuska R, "Fuzzy relational classifier trained by fuzzy clustering," *IEEE Trans Syst Man Cybern* vol.29, no.5, pp.619-625, 1999.
[31] Richard J. Hathaway and James C. Bezdek, "NERF c-means: Non-Euclidean relational fuzzy clustering," *Pattern Recognition*, Vol. 27, No. 3, pp. 429-437, 1994.

Chuanxu Wang received his B.Sc. in electronic engineering in 1990, M.Sc. in industry automation in 2000, both from Petroleum University, China, and PhD in industry automation in 2007 from The University of Ocean University, China. From Dec. 2008 to Oct. to Apr. 2009, he was a visiting researcher at AMRL at Wollongong University Australia. He is currently an associate professor in Informatics Institute, Qingdao University of Science and Technology, China. His research interests include signal processing and pattern recognition. He has published over 50 papers in refereed international conferences and journals.