Capturing Human Motion based on Modified Hidden Markov Model in Multi-View Image Sequences

Liu Yanan
Bei Jing Normal University, Zhuhai, 519087
Email: Yanan0131@sina.com

Jia Lian Kun and Yu Wen Yu
Harbin Institute of Technology at Weihai, Weihai, 264209; The Inner Mongolia Autonomous Region Sports Bureau, Huhehaote, 010030

Abstract—Human motion capturing is of great importance in video information retrieval, hence, in this paper, we propose a novel approach to effectively capturing human motions based on modified hidden markov model from multi-view image sequences. Firstly, the structure of the human skeleton model is illustrated, which is extended from skeleton root and spine root, and this skeleton consists of right leg, left leg and spine. Secondly, our proposed human motion capturing system is made up of data training module and human motion capturing module. In the data training module, multi-views motion information is extracted from a human motion database, and feature database of human motion capturing is constructed through combining multi-views motions. In the human motion capturing module, results of motion capturing can be achieved through motion classification based on a modified hidden markov model. Thirdly, the modified hidden markov model is designed by utilizing the fuzzy measure, fuzzy integer, and fuzzy intersection operator through a scaling process. Finally, a standard motion capture dataset- MPI08_Database is utilized to make performance evaluation. Compared with the existing methods, the proposed approach can effectively capture human motions with high precision.

Index Terms—Hidden Markov Model; Human Motion; Multi-View Image Sequences; Motion Classification; Confusion Matrix

I. INTRODUCTION

Human motion capturing refers to the process in which the configuration of body parts is estimated from sensor input. When poses are estimated over time, the term human motion analysis is used. Traditionally, motion capture systems require that markers are attached to the body. These systems have two major drawbacks: they are obtrusive and expensive. Many applications, especially in surveillance and Human–Computer Interaction, would benefit from a solution that is markerless. Vision-based human motion capture systems attempt to provide such a solution through cameras as sensors [1].

Automatic capture and analysis of human motion is a highly active research area due both to the number of potential applications and its inherent complexity. This research field includes a great number of difficult problems such as inferring the pose and motion of a highly articulated and self-occluding non-rigid 3D object from images. This complexity makes the research field challenging from a purely academic point of view. From an application perspective computer vision-based methods often provide the only non-invasive solution making it very attractive [2].

In recent years, automatic human motion capture has been widely researched. Traditional methods can be divided into two modes which are marker-based mode and markerless mode. For the first mode, researchers adopt a set of optically distinguishable markers placed in some landmarks of the body and are able to produce highly accurate results, mostly employed by the cinema and medical industries. However, the hardware required to acquire these markers constrains the analysis to a setup scenario and the marker placement might be intrusive and/or uncomfortable for users [3] [4].

The development of motion capture systems has exerted influence on the speedy generation of motion for 3 dimensions characters for animation and clinical studies in gate analysis. Current motion capture devices capture human motion at high frequency and the motion can be retargeted to computer characters for realistic movements [5]. However, motion generation often involves long motion capture sessions and careful editing in order to adapt and implement a particular behavior of a character. The captured motion is usually often archived and annotated for future use. The motion database is probably the most important asset in a production studio. Recorded motions are frequently recalled from the database and new motions are reconstructed from the given clips.

As is shown in Fig. 1, some examples for human motion capture are illustrated, and it can be see that key points should be determined to present the human body.

In this paper, in order to enhance the precision of human motion capturing, a modified hidden markov model is utilized. The Hidden Markov Model is a popular
statistical tool for modeling a wide range of time series data. Particularly, the Hidden Markov Model is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence. The hidden Markov models have been found to be very useful for a wide spectrum of applications in ecology, cryptanalysis, image understanding, speech, handwriting recognition, video/image processing [6-9].

The main innovations of this paper lie in the following aspects:

1) Human motions are detected using motion classification based on the proposed modified hidden Markov model.

2) The relation between the modified hidden Markov model and the standard hidden Markov model is made, in which the transitional parameters are permitted to vary with time.

3) The proposed modified hidden markov model achieves naturally and dynamically as a byproduct of the nonlinear aggregation of information using the fuzzy integral.

The rest of the paper is organized as the following sections. Section 2 introduces the related works. Section 3 illustrates the proposed scheme for capturing human motion based on modified hidden markov model. In section 4, experiments are conducted to make performance evaluation with comparison to other existing methods. Finally, we conclude the whole paper in section 5.

II. RELATED WORKS

Human motion capturing has been a hot research field in recent years, and a series of methods has been proposed. In the following section, we will survey the related works.

Yang et al. proposed a novel representation of articulated human actions and gestures and facial expressions. The main goals of the proposed approach are: 1) to enable recognition only by a few examples, and 2) meaningful organization of unlabeled datasets by unsupervised clustering. The proposed representation is obtained by automatically discovering high-level sub-actions or motion primitives, by hierarchical clustering of observed optical flow in four-dimensional, spatial, and motion flow space [10].

Liu et al. proposed a novel method for human action recognition based on boosted key-frame selection and correlated pyramidal motion feature representations. Instead of using an unsupervised method to detect interest points, a Pyramidal Motion Feature (PMF), which combines optical flow with a biologically inspired feature, is extracted from each frame of a video sequence. The AdaBoost learning algorithm is then applied to select the most discriminative frames from a large feature pool. In this way, the authors obtain the top-ranked boosted frames of each video sequence as the key frames which carry the most representative motion information [11].

Silverberg et al. studied the highly energized collective motion of attendees at heavy metal concerts, and then they found these extreme social gatherings generate similarly extreme behaviors: a disordered gaslike state called a mosh pit and an ordered vortexlike state called a circle pit. Both phenomena are reproduced in flocking simulations demonstrating that human collective behavior is consistent with the predictions of simplified models [12].

Hutchinson et al. investigated the existence, nature and extent of the binocular advantage for encoding second-order global motion. Motion coherence thresholds (79.4% correct) were assessed for determining the direction of radial, rotational and translational second-order motion trajectories as a function of local element modulation depth (contrast) under monocular and binocular viewing conditions. Furthermore, the authors found a binocular advantage for second-order global motion processing for all motion types [13].

Wei et al. explored the adsorption and thermal motion of transacting activator of transcription (TAT) peptide-modified nanocargo on a model lipid bilayer in the nonelectrostatic domain. The authors found that the lateral and rotational motion of TAT peptide-modified nanocargo could be effectively suppressed on the surface of neutral lipid membrane, a feature that cannot be explained by existing hypotheses [14].

Mather et al. indicated that the influence of form signal motion processing is more extensive than previously thought. First, the salience and apparent direction of moving lines depends on how the local orientation and direction of motion combine to match the receptive field properties of motion selective neurons. Second, orientation signals generated by “motion-streaks” influence motion processing: motion sensitivity, apparent direction and adaptation are affected by simultaneously present orientation signals. Third, form signals generated by human body shape influence logical motion processing, as revealed by studies using point-light motion stimuli [15].

Cimen et al. presented a new approach to identify the directly or indirectly related descriptors to emotion classification in human motion and conducting a comprehensive analysis of these descriptors (features) that fall into three different categories: posture descriptors, dynamic descriptors, and frequency-based descriptors in order to measure their performance with respect to predicting the affective state of an input motion [16].

Qi et al. proposed a high level semantic feature in a low dimensional space to represent the essential characteristic of different motion classes. On the basis of
the statistic training of Gauss Mixture Model, this feature can effectively achieve motion matching on both global clip level and local frame level [17].

Nascimento et al. proposed a new way of modeling trajectories, based on a mixture of parametric motion vector fields that depend on a small number of parameters. Switching among these fields follows a probabilistic mechanism, characterized by a field of stochastic matrices. This approach allowed representing a wide variety of trajectories and modeling space-dependent behaviors without using global nonlinear dynamical models [18].

Roudposhti et al. presented an approach for modeling human interactions based on existent relationship characteristics between body parts motions and environmental parameters. Human interactions properly cannot be identified without knowing the relations between the objects such as human-robot and human-human. During any human interaction, there are many relations between human body parts and others. Moreover, in this paper, a general model to analyze human interactions based on the existent relationships is presented as well. To study human motion properties, Laban Movement Analysis (LMA), a well-known human motion descriptor is used [19].

Avizzano et al. illustrated a mathematical model to generate human-like motion trajectories in space. The authors utilized linear regression in a latent space to find the model parameters from a set of demonstration examples. The learning procedure requires a relevant set of similar examples. The apprehended models encode both the typical shapes of motion and their variability towards specific boundary conditions. Particularly, the authors developed a robust tool to gather the model from examples, and to achieve real-time trajectory adaptation [20].

III. THE PROPOSED SCHEME

A. Illustration of the Three-Dimensional Human Model

Three-dimensional human model refers to a local coordinate system, in which the bones obey the Parent-Child relationship as shown in Fig. 2. The upper node of the human skeleton model is named skeleton root, which can connect with the spine root. Particularly, the spine root can connect legs with spine. In the given human skeleton model, the remainder of the human body can be extended from the leg, including thigh, shin, foot and toes. Moreover, spine is an important part in the three-dimensional human model, which can connect human’s upper body, such as shoulder, arm, hand, neck and head.

Based on the above human skeleton model, three-dimensional virtual human model can be designed. Afterwards, the human motion capture problem can be converted into a convex problem and then be simplified utilizing a hierarchical geometrical solver. In the given human skeleton model, parameters are set for the hierarchical procedure starting from the hip and all the way up to the head. Particularly, this model integrates the image data to overcome the problem of being short of information about the model parameters.

B. Capturing Human Motion Based on Modified Hidden Markov Model

Using the above three-dimensional human model, the method of human motion capturing can be illustrated. As is shown in Fig. 3, framework of the human motion capturing system based on modified hidden markov model is given.

The proposed human motion capturing system consists of two modules. Module 1 implements data training, in which a human motion database is constructed. Then, multi-views motion information is processed. Afterwards, feature database of human motion capturing is constructed. Then, feature database of human motion capturing is obtained by integrating multi-views motion and extracting motion
descriptor from the proposed human motion database. In module 2, the sample video with multi-view human motion is used as the input data, and then multi-view image sequences are extracted. Based on the above works, a modified hidden Markov model is proposed. Finally, results of human motion capturing can be obtained through motion classification based on the proposed modified hidden Markov model.

Considering there is a computational difficulty in traditional hidden Markov model. Therefore, in this scheme, we will modify the traditional hidden Markov model by introducing the fuzzy measure, fuzzy integer, and fuzzy intersection operator through a scaling process.

Therefore, we extended the standard hidden Markov model by combining the fuzzy Mathematics theory in it. The formalized descriptor of the modified hidden Markov model is represented as \( \lambda = (A, B, \pi) \). Furthermore, for each state \( S \), a symbol fuzzy measure \( b_i(\cdot) \) is defined on the space of the observation vectors \( \Omega \). For the time slot \( s \in [1,T] \), values of the symbol fuzzy densities are defined as follows.

\[
b_i(Q_s), b_2(Q_s), \ldots, b_M(Q_s)
\]

where \( M \) refers to a fuzzy set on the set which have \( M \) states. Hence, \( T \) different fuzzy sets on the set of \( M \) states are built. Therefore, \( b_i(Q_s) \) could be used as the membership value of \( Q_s \) in the state \( S \). Particularly, the interpretation for the transition fuzzy density lies in that the parameter \( a_{ij} \) can measure the grade of certainty of the statement which will reach the state \( S_j \) at time \( t+1 \).

Afterwards, the set of \( K \) training observation sequences is defined as follows.

\[
Q = \{Q^1, Q^2, \ldots, Q^K\}
\]

where \( Q^k = \{Q^k_1, Q^k_2, \ldots, Q^k_T\} \) refers to the \( k \)th observation sequence. Next, the parameters of the proposed modified hidden Markov model can be obtained by maximizing the following equation as follows.

\[
P(Q|\lambda) = \prod_{k=1}^{K} P(Q^k|\lambda) = \prod_{k=1}^{K} P^k
\]

Afterwards, the modified re-estimation equation for the transition fuzzy measures is represented as follows.

\[
a_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T-1} a_{ij}(i) \cdot a_{ij}(i, j) \cdot b_j(Q^k_{t+1}) \cdot b_i(Q^k_t)}{\sum_{k=1}^{K} \sum_{t=1}^{T-1} a_{ij}(i) \cdot a_{ij}(i, j) \cdot b_j(Q^k_{t+1}) \cdot b_i(Q^k_t)}
\]

where \( b_j(\cdot) \) refers to the conditional fuzzy measure on \( \Omega \) with respect to the state \( S_j \), and \( a_{ij} \) denotes the transition fuzzy density.

Based on the modified Hidden Markov model, human motion could be captured through a motion classification process. The multi-views image sequences should be classified in advance. In order to recognize the motion of the input multi-views image sequences, the most important problem is how to seek the best state sequence. Moreover, the observation sequence is represented as \( U = \{U_1, U_2, \ldots, U_T\} \), and a model is defined as \( \psi = (A, B, \Theta) \). The time vector is shown as follows.

\[
q^k = (q_{1,1}, q_{2,2}, \ldots, q_{T, T})
\]

where \( q_i \) refers to the actual state at time \( i \). Given the above modified hidden Markov model \( \psi \) and the observation set \( U_i(t) \), the parameters of the proposed model can be obtained by maximizing the likelihood function which is used for motion classification. Hence, the model \( \psi \) can be optimized by the following equation.

\[
\psi = \arg \max_{\psi \in \text{setences in database}} P(U|\psi)
\]

where \( U \) denotes the unknown feature vector sequence of a sample motion, and \( \psi \) refers a hidden markov model from the set of all given motions. Afterwards, the classifier can be implemented by seek the model \( \psi \) with the highest conditional probability.

Next, the observation density is calculated by the following equation.

\[
b_1(u) = \sum_{k=1}^{K} c_{\beta} \cdot G(u, \mu_\beta, \sum_j k)
\]

where \( u \) refers to the observation vector, and \( c_\beta \) represents the mixture coefficient for the \( \beta \)th mixture in the state \( j \). \( G \) denotes the Gaussian, and it can be computed as follows.

\[
G(u, \mu_\beta, \sum_j k) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (u - \mu_\beta)^T \Sigma^{-1} (u - \mu_\beta)}
\]

The number of the modified hidden Markov model states is determined by the average motion signal length, the complexity and the value of the variability. Then, the number of key features for the motion sequence is selected as the number of the markov states.

IV. EXPERIMENT

In this section, performance evaluation of the proposed method is given using a standard motion capture dataset-MPI08_Database, which is designed by multi-sensor fusion for 3D full-body human motion capture [21] [22]. The proposed dataset MPI08_Database consists of five parts, including: (a) sequences: multi-view sequences obtained from 8 calibrated cameras, (b) silhouettes: binary segmented images obtained with chroma-keying, (c) meshes: 3D laser scans for each of the four actors in the dataset and also the registered meshes with inserted skeleton, (d) projection matrices: one for each of the 8 cameras, (e) Orientation data: raw and calibrated and
sensor orientation data. Some of the scenes from MPI08 Database are shown in the Fig. 4.

Utilizing the MPI08 Database, we test the performance of the proposed human motion capturing algorithm for four different postures, which are 1) P1: Cartwheel, 2) P2: Jumping jack, P3: Skiing, 4) P4: Rotating both arms and 4) P5: Kicking. Other three related methods are compared with the proposed algorithm, which are LSVM, SSTF, and VF. Firstly, confusion matrices for the above three methods and the proposed algorithm is shown in Fig. 5.

Afterwards, we will give the overall performance of the above four methods in Fig. 6.

In the following section, we will show the recognition accuracy by the proposed algorithm under each single view image and multi-view images (shown in Table.1).

As is shown in Table.1, using the multi-view image sequences, the proposed algorithm can obviously promote the recognition accuracy of human motion capturing.

From the above experimental results, it can be seen that the proposed scheme is superior to other schemes. The main reasons lie in the following aspects:

The LSVM method implements the motion recognition by combining local features with SVM. In this method, the authors demonstrate how such features can be used for recognizing complex motion patterns. Next, LSVM constructs video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. Particularly, to represent motion patterns, this method utilize local space-time features which can be considered as primitive events corresponding to moving two-dimensional image structures at moments of non-constant motion.

2) In SSTF, a new spatio-temporal interest point detector is presented, and a number of cuboid descriptors are analyzed. SSTF tries to establish the link between the domains of behavior recognition and object recognition, creating the potential to bring in a range of established techniques from the spatial domain to that of behavior recognition. However, a dynamic model should be integrated into this method.
In this paper, a novel human motion capturing algorithm is proposed. The proposed human motion capturing system is made up of two modules, which are 1) data training module and 2) human motion capturing module. In module 1, multi-views motion information is obtained through motion classification based on the fuzzy integral.

Particularly, we modify the standard hidden Markov model in that we construct a relation between the modified hidden Markov model and the standard hidden Markov model. This allows for the transitional parameters to be modified naturally and dynamically as a byproduct of the recognition accuracy, an appearance model should be added into the VF method.

4) Compared with the above three approaches, the proposed algorithm has the following three advantages:

a) The main innovation of the proposed algorithm lies in that we construct a relation between the modified hidden Markov model and the standard hidden Markov model in which the transitional parameters are allowed to vary with time.

b) The proposed modified hidden markov model achieves naturally and dynamically as a byproduct of the nonlinear aggregation of information using the fuzzy integral.

c) Results of human motion capturing of our approach are obtained through motion classification based on the proposed modified hidden markov model.

V. CONCLUSION

In this paper, a novel human motion capturing algorithm is proposed. The proposed human motion capturing system is made up of two modules, which are 1) data training module and 2) human motion capturing module. In module 1, multi-views motion information is extracted from a human motion database, and feature database of human motion capturing is constructed through combining multi-views motions. In module 2, human motions can be captured through motion classification based on a modified hidden markov model. Particularly, we modify the standard hidden markov model based on the fuzzy measure, fuzzy integral, and fuzzy intersection operator.

REFERENCES


