A Real-time Human Motion Recognition System Using Topic Model and SVM

Jie Li, Ting Ma*, *Member, IEEE*, Xiaorong Zhou, Yingke Liu, Shuo Cheng, Chenfei Ye, Yutong Wang

Abstract—**Human motion recognition is a challenging task, especially when motion capture data is huge. Existing approaches for this task focused mainly on how to extract features from motion capture data to achieve high recognition performance. However, due to the presence of redundant features and the high dimensionality of data, these approaches may not achieve the optimal performance. In order to rapidly and accurately recognize human motion, we present a novel method based on topic model and SVM. It consists of first extracting informative angle features at each frame of motion sequence, which is helpful to eliminate redundant information. Then, significant motion frames are transformed into motion-words to form motion sequences. A motion sequence is a combination of motion-words, which are obtained by extracting key poses and applying hierarchical clustering. Then, topic model are applied to obtain the underlying topic distribution to perform motion recognition. We studied 9 common motions for 13 subjects. The results show the best recognition accuracy achieves 98.41% which outperforms state-of-the-art methods.**

I. Introduction

Human motion recognition has been an active field of research for more than a decade [1]. It has applications in various fields, such as fitness tracking, health monitoring, fall detection, rehabilitation medicine, virtual reality and smart home [2], [3]. Human motion recognition has also been used in interdisciplinary research areas, for example by combining emotion recognition [4] and psychiatric disease diagnosis [5].

Existing approaches to recognize human motion fall into two main classes [6]. The first one is approaches based on computer vision, which are inspired by how humans detect motions. This approach has two important disadvantages: (1) it does not provide kinematic information, and (2) it is easily affected by background noise. The other class is based on inertial sensors. It is portable, precise and can be used in real-time. In particular, wearable inertial sensors have attracted widespread attention from researchers in recent years because of their portability and availability.

Motion capture data, obtained by wearable inertial sensors, is high dimensional and sequential. Many researchers have attempted to extract appropriate features from this data to describe motions. Zhu et al. [7] found that the combinations of geometric features could more accurately describe information than any single feature. Hence, they adopted distance and angle features to represent motions. But this method ignored that motion intensity can vary greatly for different individuals. Leightley et al. [8] used time-frequency features of skeletal joints as feature vector. Zhu et al. [9] utilized the normalized relative orientations of human joints as skeletal features. In [10], Gowayyed et al. used histograms of oriented displacements (HOD) to represent the 3D trajectories of body joints to recognize human motions. To provide some qualitative insights in how actions are related to the motions of the body, Ofli et al. [11] used sequences of the most informative joints (SMIJ) to describe motions. They picked out some skeletal joints that are considered as the most informative for representing motion sequence. However, they also faced the problem that many features are redundant and that motion data is high dimensional. This result in decreased accuracy and increased computational complexity for motion recognition approaches.

In this paper, to rapidly and accurately recognize human motion, we present a real-time human motion recognition system based on topic model and SVM. Firstly, we extract meaningful angle features at each frame of a motion using coordinate transformation. Basic units are created by key pose extraction and hierarchical clustering, which also called motion-words. Then, each frame is replaced by motion-word representing the most similar pose. Thus, motion capture data is transformed into a text-like document. Next, the proposed approach extracts the topic distribution of training motion using topic model. Finally, SVM is trained to recognize motions using the topic distribution.

The main contributions of this work are two folds. First, meaningful angle features are extracted from motions. This method not only eliminates redundant information in motion capture data, but also efficiently describes motions. Second, a novel approach proposed which applies a topic model to obtain the patterns of motions and SVM is trained to recognize motions. Experimental results show that this method provides excellent classification performance and is efficient.

The remainder of this paper is organized as follows: Section II describes the proposed real-time human motion system. The method is evaluated by extensive experiments, described in Section III. Finally, Section IV provides a discussion and draws a conclusion.

II. Methodology

In this Section, we give a detail description of our system. The flowchart of real-time human motion recognition system is shown in Fig. 1.

Research supported by the Basic Research Foundation of Shenzhen Science and Technology Program (JCYJ20150403161923510), Basic Research Foundation of Shenzhen Science and Technology Program Kep Project Track (JCYJ20160509162237418), the Key Cultivating Project of Harbin Institute of Technology, and High-end Talent Oversea Returnees Foundation of Shenzhen (KQC201109020052A).

Jie Li, Ting Ma, Xiaorong Zhou, Yingke Liu, Shuo Cheng, Chenfei Ye and Yutong Wang are with the Department of Electronic & Information Engineering, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China. (Corresponding author: Ting Ma. Phone: +86-755-26033608; fax: +86-755-26033608; e-mail: tmahit@outlook.com).

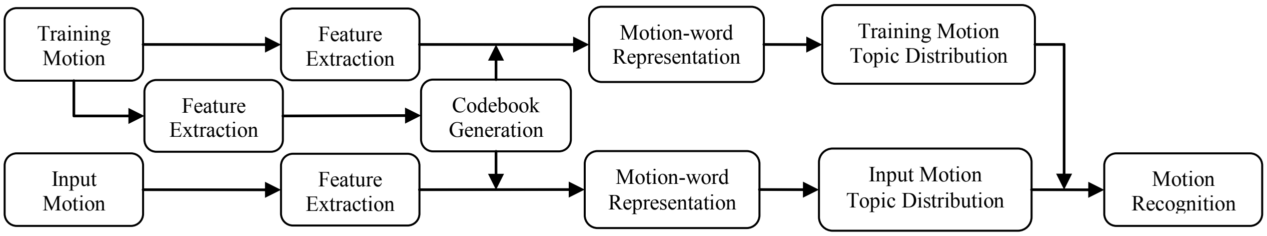

Figure 1. A flowchart of real-time human motion recognition system

A. Data Collection

In this study, we have investigated 9 motions, including bowing, walking, running, sitting, lying, squatting, tripping, walking up the stairs and walking down the stairs. We recruited 13 volunteers to collect the motion data at 120fps using Noitom Perception Neuron. Perception Neuron composed of a 3-axis gyroscope, 3-axis accelerometer and 3-axis magnetometer. And it operates with 18 individual sensors that can be placed on the body. Each motion is performed by each volunteer for five times in each sequence. The length of each motion sequence is about 5 seconds. Hence, we obtained 685 motion sequences in total.

B. Feature Extraction

Human motion is typically addressed by the movement of the different body parts [12], [13], which include the distance between the hand and hip, and the angle of various body parts. For example, some human daily motions, including walking, running and sitting, can be described by $Angle_1$, $Angle_2$ and $Angle_3$ in Table I. Since the Euler angles obtained by the wearable sensor are unstable and noisy [15], the motion capture data are translated to global coordinate system $p_{j,n}$ ($x_{j,n}$, $y_{j,n}$, $z_{j,n}$) at nth frame for joint j by [14], [19]. We assume that the 3D coordinates at nth frame were $p_{5,n}$ ($x_{5,n}$, $y_{5,n}$, $z_{5,n}$) for the left thigh, $p_{6,n}$ ($x_{6,n}$, $y_{6,n}$, $z_{6,n}$) for the left calf, $p_{0,n}$ ($x_{6,n}$, $y_{6,n}$, $z_{6,n}$) for the hip. At nth frame $Angle_2$ can be calculated by (1). Other angles are the same as that. So we extract 6 angle features in Table I at each frame of the motion. Then, feature scaling used to bring all angle features into the range [0, 1].

$$Angle_{2,n} = \frac{arc\cos\left(\left(p_{0,n} - p_{5,n}\right)\cdot\left(p_{6,n} - p_{5,n}\right)\right)}{\left|p_{0,n} - p_{5,n}\right|\cdot\left|p_{6,n} - p_{5,n}\right|} \tag{1}$$

TABLE I. ANGLE FEATURE

Name	Angle Range (°)	Description
$Angle_1$	0 ~ 90	The angle between the waist and the ground anchor.
$Angle_2$	0 ~ 180	The angle between the left thigh and the left calf.
$Angle_3$	0 ~ 180	The angle between the right thigh and the right calf.
$Angle_4$	0 ~ 90	The angle between the left thigh and right thigh forward.
$Angle_5$	0 ~ 180	The angle between the left upper arm and left lower arm.
$Angle_6$	0 ~ 180	The angle between the right upper arm and right lower arm.

C. Codebook Generation

Each human motion sequence can be described as a natural language by motion-words. In this paper, the motion-word is obtained by extracting key pose and hierarchical clustering. We regard the first frame and the last frame as key poses. To find next key pose, we calculate the similarity between the previous key pose and each frame remaining in the sequence until the similarity is less than threshold. The similarity between A and B is given in (2). This procedure is repeated until the last frame is traversed. Through these steps, we can obtain all key poses from training motions. Then we use hierarchical clustering to merge similar poses. The cluster center is motion-word, and the set of all motion-words is codebook.

$$Similarity(A,B) = \frac{A \cdot B}{|A|\cdot|B|} \tag{2}$$

D. Motion-word Representation

As a document consists of many words, human motions can be seen as the combinations of motion-words. We replace each frame in motion sequence with the most similar pose in codebook. Since the raw motion sequence is large and redundant, it would cost a lot of time to represent the raw sequence by motion-words. Therefore, we need to extract the significant key poses from original sequence beforehand. The step is as same as the codebook generation.

Assuming that the motion sequence is $\{S(n)|n=1,\ldots,N\}$, where N is the length of this sequence, $S(n)$ is the nth frame in the motion sequence. The codebook is denoted as $C=\{c_1,c_2,\ldots,c_K\}$, where K is the size of codebook. We use motion-words to represent the original motion sequence as $W=\{w_1,\ldots,w_n\}$ by (3).

$$w_n = \arg\max_k Similarity(S(n), c_k) \tag{3}$$

In (3), w_n is a motion-word corresponding to the nth frame in the motion sequence.

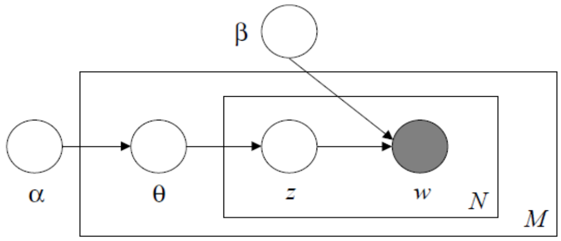

Figure 2. Graphical representation of LDA model. The larger rectangle denotes documents, while the smaller rectangle denotes the chosen topics and words within the document.

E. Topic Modeling

Latent Dirichlet allocation (LDA) is one of the most classic topic models that can discover latent topics of human motion. The basic idea is that each topic is seen as a distribution over words, and each document is regarded as a distribution over latent topics [16]. Figure 2 shows a graphical representation of LDA model. As a matter of convenience, we define some following notations. M is the number of documents and N_m is the number of words in the mth documents. Two parameters α and β are the Dirichlet prior vectors for the topics in per-documents and words in per-topics, respectively. It's worth noting that two hyperparameters α and β are related to θ and φ [15]. If we want to generate corpus C that contains M documents, the generative process is as following:

1) For each document i ($1 \le i \le M$), choose $\theta_i \sim Dirichlet(\alpha)$; θ_i is the topic distribution for document i.

2) For each word j ($1 \le j \le N_m$) in the document i:
   a) Pick out a topic $z_{i,j} \sim Multinomial(\theta_i)$;
   b) Pick out a word $w_{i,j} \sim Multinomial(\varphi_{z_{i,j}})$.

After motion sequence is transformed into a text-like document, topic model can be used to construct the pattern of motion. We regard different types of motions as documents of different corpus. For the same type of motions, the same label is attached to each frame. At last, we can discovery the latent topic using topic model and match the topic with corresponding motion.

Topic model considers not only the semantic relativity among words but also topic distributions in document. We connect all training motion sequences, which are represented by motion-words, and form a one-dimensional vector as the input of topic model. By counting the frequency of each word occurring in the document, we can obtain the word distribution in the document. Then we get the topic distribution by using LDA model. Through these steps, we have completed all work in training phase. The topic distribution of input motion sequence can be calculated by (4), which reflects the share of various key poses in motion.

$$\theta_t = \sum_{c=1}^{K} \frac{S_c \theta_t \varphi_{t,c}}{\sum_{j=1}^{T} \theta_j \varphi_{j,c}} \tag{4}$$

In (4), S_c is the total number of motion word c occurring in the input sequence, θ_t is the probability of tth topic in training phase, $\varphi_{t,c}$ is the probability of motion word c in topic t.

F. Motion Classification

As SVM is robust and effective in practical issues, we choose SVM to solve human motion recognition problem in this paper. Kernel function is the key part of SVM, which is used to measure the similarity of two data objects. Radial basis function (RBF) can not only map samples into a higher dimensional space, but can also control the relationship between class labels and attributes when they are non-linear [17]. Thus, we choose RBF as the kernel function. RBF is defined as (5), where x and y are training label and testing label respectively, γ is the kernel parameter.

$$K(x,y) = \exp\left(-\gamma \|x-y\|^2\right), \gamma > 0 \tag{5}$$

In previous subsection, we obtained the topic distribution of training motion. There is significant difference among topic distribution of different kind of motions. Hence, the pattern of motion sequence can be represented as topic distribution. To train the classifier, the topic distribution for each motion sequence is stamped with its class manually. Then we compute the topic distribution of novel motion sequence by (4), and get the recognition result through SVM.

III. Result

The presented algorithm has been implemented by self-developed software in Matlab. We evaluated our proposed human motion recognition system by leave-one-out cross-validation with one subject for testing and the rest of subjects for training. Fig.3 shows the recognition performance in different size of codebook. It can be found that the recognition performance increases alone with the increasing size. When the size is larger than 30, the best recognition performance is achieved with no noticeable change. Thus, it's important for topic model to find an appropriate size of codebook. Excess motion-words would lead to the increase of computational complexity.

In LDA model, we need to initialize three parameters: the number of topics T, two hyperparameters α and β. We can refer to [18] to set two hyperparameters. Here we investigate the effect of the number of topics T, as shown in Fig.4. We can find that the changes of accuracy are very dramatic at the smaller topic numbers. Then they are becoming flat as the number increases. It is probably because some informative underlying topics are gradually discovered as T increase. When T is enough large, all underlying topics are found.

We also compared our method with several state-of-the-art motion feature descriptors [8], [10], [11], as shown in Table II. The time-frequency feature is represented by 225 values (15 joints × 3 for three projections × 5 features). These features include mean, variance, skewness, kurtosis and energy. We use 4-bin HOD and 3-level temporal pyramid as another motion feature descriptor, which achieves good performance in our study. For SMIJ features, the best recognition result is obtained by setting the number of top-ranking joints to be 6 and the number of segments to be 60. Our proposed method choose optimal parameters (e.g., T=20, α=50/T, β=0.01, K=30). We use SVM to evaluate above methods. In Table II, we can find that our method is superior to other state-of-the-art methods in higher accuracy. It means that the proposed method can effectively improve the recognition performance.

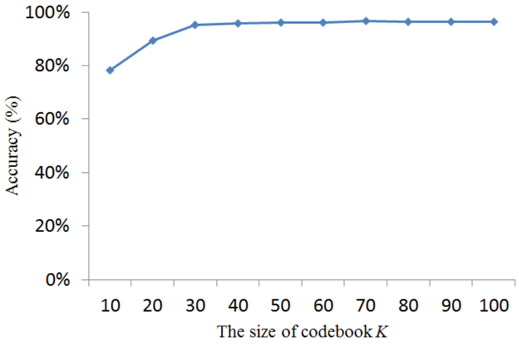

Figure 3. The influence of the size of codebook K when T=20.

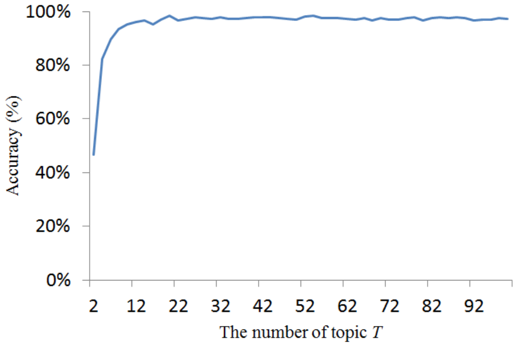

Figure 4. The influence of the number of topics *T* when *K*=30.

TABLE II. COMPARISON OF OPTIMAL RECOGNITION PERFORMANCE OF THE STATE-OF-THE-ART METHODS AND OUR PROPOSED METHOD

Method	Accuracy (%)
Our proposed method	**98.41**
Time-Frequency Feature [8]	88.57
HOD [10]	91.48
SMIJ [11]	93.54

IV. DISCUSSION

In this paper, we presented a real-time human motion recognition system, which consists of feature extraction, codebook generation, motion-word representation, topic modeling and motion recognition. Here we discussed three state-of-the-art methods, as shown in Table II. Using common time-frequency feature [8] could achieve good recognition performance, while it is easy to result in "dimension curse". Gowayyed et al. [10] used HOD to describe the motion capture data. The HOD is constructed by counting the length of each displacement in an orientation angles histogram. However, the method put the focus on all skeletal data, which increased computational complexity and imported noise. Ofli et al. [11] used sequence of the most informative joints (SMIJ) as features. The joints are selected by highly interpretable measures, including the mean of joint angles, the variance of joint angles and maximum angular velocity of joints. In comparison with HOD, this method efficiently decreased computational complexity and reduced noise by picking out some important joints, so the accuracy is greater than that of HOD. However, the SMIJ features ignored the effect of motion speed.

In comparison to three methods, our approach extracted 6 most informative angle features to describe human motion in advance. It contributes to eliminate redundant information in the original motion capture data. Then we applied LDA to get the latent topic, which has more semantic information. According to the topic distribution, the pattern for each motion can be obtained accurately. Meanwhile, the selection of appropriate number of topics could degrade the computational complexity. Hence, our method achieved the best recognition performance.

This study has limitations. On the one hand, the orientation of human is hard to be determined. On the other hand, this study mainly concentrates on the movement of lower limbs. Our future work will focus on how to determine the orientation of human and extend the study into more complex motions.

ACKNOWLEDGMENT

Research Foundation of Shenzhen Science and Technology Program (JCYJ20150403161923510), Basic Research Foundation of Shenzhen Science and Technology Program Kep Project Track (JCYJ20160509162237418), the Key Cultivating Project of Harbin Institute of Technology, and High-end Talent Oversea Returnees Foundation of Shenzhen (KQC201109020052A).

REFERENCES

[1] Lara O D, and Labrador M A. A survey on human activity recognition using wearable sensors. IEEE Communications Surveys & Tutorials, 15(3), 2013.

[2] Jeffrey W L, Tony P, and Gary M W, "Applications of mobile activity recognition," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ACM, 2012, pp. 1054-1058.

[3] Benmansour A, Bouchachia A, and Feham M, "Human activity recognition in pervasive single resident smart homes: State of art." pp. 1-9.

[4] Bevilacqua V, Barone D, Cipriani F, Onghia G D, Mastrandrea G, Mastronardi G, Suma M, and Ambruoso D D, "A new tool for gestural action recognition to support decisions in emotional framework." pp. 184-191.

[5] Tacconi D, Mayora O, Lukowicz P, Arnrich B, Setz C, Troster G, and Haring C, "Activity and emotion recognition to support early diagnosis of psychiatric diseases." pp. 100-102.

[6] Daniel V, Rolf A, Giovanni V, John B, Markus G, Wojciech M, and Jovan P, "Practical motion capture in everyday surroundings," *ACM SIGGRAPH 2007 papers*, ACM, 2007, p. 35.

[7] Mingyang Z, Huaijiang S, Rongyi L, and Bin L. Human motion retrieval using topic model. Comput. Animat. Virtual Worlds 1546-4261, 23(5), 469-476, 2012.

[8] Leightley D, Darby J, Li B, McPhee J S, and Yap M H, "Human Activity Recognition for Physical Rehabilitation." pp. 261-266.

[9] Zhu G, Zhang L, Shen P, Song J, Zhi L, and Yi K, "Human action recognition using key poses and atomic motions." pp. 1209-1214.

[10] Mohammad A G, Marwan T, Mohamed E H, and Motaz E-S, "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press, 2013, pp. 1351-1357.

[11] Ofli F, Chaudhry R, Kurillo G, Vidal R, and Bajcsy R, "Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition." pp. 8-13.

[12] Aggarwal J K, and Cai Q, "Human motion analysis: a review." pp. 90-102.

[13] Bharatkumar A G, Daigle K E, Pandy M G, Cai Q, and Aggarwal J K, "Lower limb kinematics of human walking with the medial axis transformation." pp. 70-76.

[14] Cai G, Chen B M, and Lee T H, "Coordinate Systems and Transformations," *Unmanned Rotorcraft Systems*, pp. 23-34, London: Springer London, 2011.

[15] Lan R, and Sun H. Automated human motion segmentation via motion regularities. The Visual Computer, 31(1), 35-53, 2015.

[16] Blei D M, Ng A Y, and Jordan M I. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022, 2003.

[17] Hsu C-W, and Chang C-C, "Chih-Jen lin,"A practical Guide to Support Vector Classification,"", 2010.

[18] Heinrich G. Parameter estimation for text analysis. University of Leipzig, Tech. Rep, 2008.

[19] Zhang S, Dong H, Zhang Z, and Wu Q, "The role of control based on three-dimensional human model." pp. 802-807.