Real-Time 3D Motion Recognition of Skeleton Animation Data Stream

Jianchao Lv and Shuangjiu Xiao

***Abstract*—In this paper, a method for real-time 3D motion recognition based on a hierarchical recognition framework is presented. To facilitate the recognition process, motions are divided into three levels by duration and complexity. SVD (Singular Value Decomposition) is used to extract the feature vector of each motion matrix, and SVM(Support Vector Machine) is utilized to do the training and classification of the first level of motion(sub-motion). In motion recognition process, the sequence of recognized candidate sub-motions is analyzed by HMM (Hidden Markov Model) to gain certain robustness, then we recognize the second level of motions by pattern matching in this sequence. Finally a grammar-based motion synthesization is applied using motions as semantic terms to recognize the third level of motions. Experimental results show that the proposed method has high performance in sensitivity, accuracy, specialty and efficiency.**

***Index Terms*—Motion recognition, support vector machine, singular value decomposition, grammar, classification.**

I. Introduction

Recently, motion recognition can find wide applications in virtual reality as it provides a more nature way of human-machine interaction. It depends on MoCap (motion capture system) as its data provider, inertial-based MoCap has caught more and more attentions due to its capability of providing reliable real-time motion data in an explicit way [1]. Data streams generated from these MoCaps have multiple attributes, each of which is the kinematics parameter of one sampling point or joint of a human skeleton [2]. The multi-attribute property poses certain challenges to real-time motion recognition.

- **Motions differ in duration and complexity:** they can be short and simple, they can be long and sophisticated**,** they should be handled differently.
- **Massive data to deal with:** real-time motion streams have thousands of rows of multi-attribute data, which need to be reduced to data of small scales while still representative of previous data.
- **Spatio-temporal variations of similar motions:** as they may be of different durations, trajectories and scales, etc. Similar motions doesn't necessarily result in similar data, the effect of variations need to be eliminated.

Motion recognition algorithms can be roughly grouped into four classes. First, the most intuitive way is template matching, usually there is a distance metric to make pair wise comparison. Dynamic programming and DTW (dynamic time warping) and their variations are often used in similarity comparison. Ref. [3] created a citation structure called match web by using DTW, the structure is used for recognize similar motions. Ref [4] used CDP (Continuous Dynamic Planning) to locate and recognize gestures from continuous gesture sequences. Although DTW can take care of the time sequence alignment, its recognition effect is still affected by the spacial variations of two similar motions. Second, methods based on classification bring machine learning algorithms to this field. Michalis Taptis applies logistic regression to evaluate how well the dancer's performance matches with the canonical model [5] and obtains good results, however, motion curves, which are its proposed features, are not good abstraction but rather another way of presentation of motion data. Third, methods based on statistical model include HMM [6]-[8], DBN (Dynamic Bayes Network) [9], etc. Well trained statistic model can unveil the implicit statistical relations in gestures or motion sequences, thus help eliminate spatio-temporal variations in recognition process. Finally, methods based on grammar are used for good recognition of sophisticated, enduring and usually interactively motions, Ref. [10] proposed a context-free grammar to recognize motion in complicated scenarios. Methods based on grammar depend on the recognition of lower level of motions, therefore, importance lies very much in the lower level of motion recognition.

In this paper, we propose a motion recognition method based on a hierarchical recognition framework as we divide motions into three levels by their durations and complexities. The system consists of an offline training module and a real-time recognition module. We recorded thousands of sub-motions (300 for each) for training, the source data is cut by a shifting window to form the sub-motion matrices, and then SVD is conducted on each matrix to extract feature vector. We utilize SVM to train the labeled feature vectors of every two motion classes, in recognition, the trained models are used to make prediction of the most likely sub-motion. The sub-motion sequence is filtered by HMM and then assembled to be recognized as motions, the motions are basic tokens of our pre-defined grammars to synthesize motions into sentences as high level of messages or controlling commands in human-machine interaction.

Key contributions of the proposed method include:

- **A hierarchical frame work:** divide motion into three different levels and a framework consists of motion training and recognition pipelines to provide appropriate methods for each level of motion.
- **Motion data abstraction and spatio-temporal variation elimination:** using SVD, one-versus-one multi-class SVM and HMM.
- **Definition of an extendable grammar:** allow the

Manuscript received July 1, 2013; revised Septemeber 5, 2013.

The authors are with Shanghai Jiaotong University/School of Software, Shanghai, P. R. China (e-mail: lvjianchao@sjtu.edu.cn, xsjiu99@cs.sjtu.edu.cn).

machine understand higher levels of human motions.

The remainder of this paper is organized as follows. Section 2 describes the proposed method in details. Section 3shows the results of our method and in Section 4, future works are discussed and conclusions are drawn.

II. Our Approach

A. Motion Data Source and Skeleton Representation

Vision-based motion capture systems often encounter problems such as occlusion of parts of the body, background noise, restrictions such as lighting or speed of movement. Moreover, in order to capture a 3D motion, multiple cameras need to set up, thus the user's moving is restricted in the cameras view.All of these disadvantages of vision-based motion capture system may hinder the motion recognition's efficiency, sensitivity and accuracy, to overcome this, we use the motion capture system based on inertial sensing.It can tracking a user's subtle motion with fast speed,high precision in a direct way, all need to do is putting on a suit with certain number of inertial sensors, each sensor is called a IMU (Inertial Measurement Unit), itcan detect current rate of acceleration with accelerators,detect changes in pitch, roll and yaw bygyroscopes and calibrate against orientation drift using magnetometers.In our experiment, we use Animazoo IGS-180-I suit, of whichthere are 18 IMUs mountedin it (see Fig. 1). Each IMUcan provide this body part's real-time rotational data in quaternions(at the rate of 50fps), all of them are connected by daisy chain wire, thus an animation skeletonis formedwith head as its referenceroot(see Fig. 1), as head skin has leastsoft tissues which can affect the IMU's alignment.

B. Motion Levels

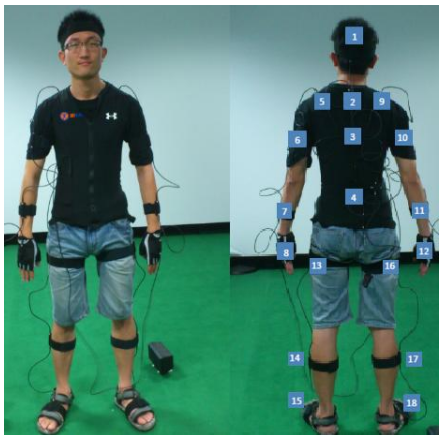

Fig. 1. Inertial MoCap Sensor Distribution.

Motions vary in durations and complexities, Aaron F. Bobick attaches importance to dividing motions into different levels in perception of motions, the division, as he argues, is necessary in making explicit both there presentational competencies and manipulations in motion recognition[11]. Most motion recognition methods focus on one level of motion or even mix up with different levels, nevertheless, the higher level of motion is based on the lower one and they should be treated with different methods, after all, there isn't a method applicable everywhere. In this paper, we divide motions into three levels:

- **L_1M:** the atomic motion (or movement), which is considered the unit of recognition by our approach, and is the sub-motion of L_2M.
- **L_2M:** the lexical motion (or activity), which is a motion that has lexical meaning and can act as the term of our grammar, for instance, the parameter motion, "Left Direction", which consists of a "lift up left hand over head" and two "leftward swing of left arm".
- **L_3M:** the **syntax** meaningful motion is a sentence with which the computer can take as the messages or controlling commands to follow in virtual reality, which offers more natural and easy way of human machine interaction.

C. Method Overview

With overcoming the aforementioned difficulties of motion recognition in mind, a hierarchical framework of a training pipeline and a recognition pipeline is proposed in order to obtain high sensitivity and accuracy in recognition of all three levels of motions. In both training and recognition pipelines, the motion matrices are generated first. MoCap provides data at the rate of 50 fps, and we define a standard L_2M's duration is of 1 second, so there are 50 frames, each frame includes 18 quaternions. Therefore, motion matrices are$n \times (s \cdot q)$matrices where n (50) denotes the number of frames, s(18) the number of sensors and q(4) the number of elements in a quaternion.

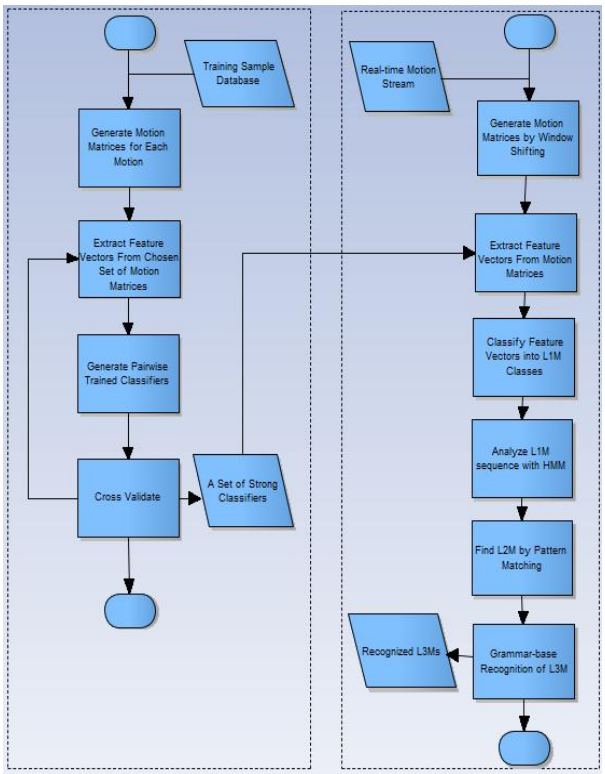

Fig. 2. The flowchart of motion training and recognition pipelines.

Feature vectors of a set of chosen motion matrices(>300) are extracted by using SVD, and we propose combining the first and second singular vectors of motion matrices to form the feature vectors, which reveal the geometry structures of them. By training these feature vectors, we get a set of pairwise classifiers, the strong classifiers is finally chosen after cross validation to assure its quality, if accuracy of

classification by cross validation is not ideal, we iterate this procedure as Fig. 2 illustrates. Finally, a set of $n(n-1)/2$ strong classifiers is settled, where n denotes the number of L_1Ms and will be used in the recognition pipeline.

D. Feature Extraction

Motion matrices contain too much redundant information for classifiers, which will seriously affect the classification accuracy if applied SVM directly. Fig. 3 shows a 3D view of a motion matrix, although it has too many elements, its structural feature can be easily seen, an intuitive thinking is extracting the geometry structure from it. SVD is a basic while powerful tool to compute the geometry objects of a large matrix by decomposing it into three matrices, where each of them represent its information in a particular aspect. The decomposition depend on the theorem[12]: Let $A \in R^{m\times n}$ (real $m \times n$matrix), then the following equation will be established:

$$\mathrm{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} \tag{1}$$

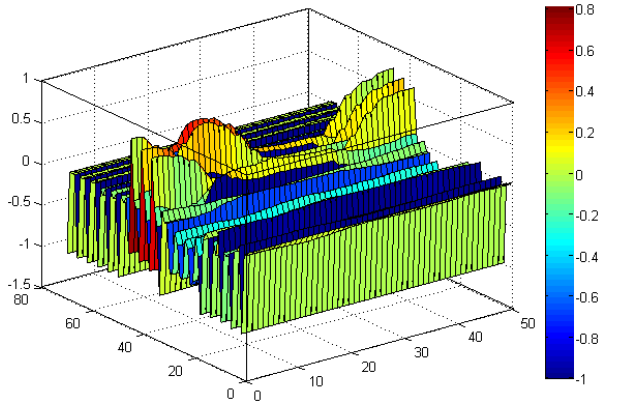

Fig. 3. 3D view of a motion matrix.

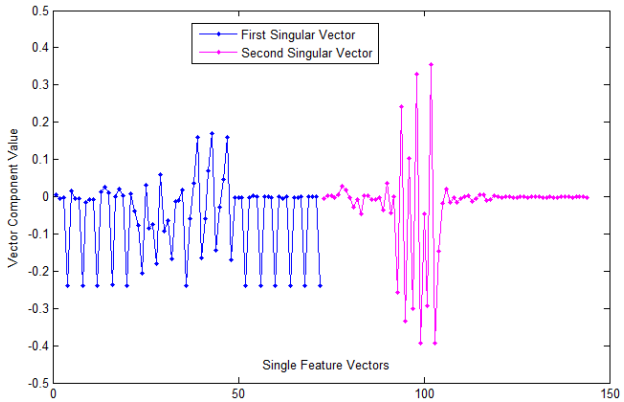

Fig. 4. The first and second singular vectors of a sub-motion.

where $U \in R^{m\times m}$ (real $m \times n$ matrix) and $V \in R^{n\times n}$ (real $n \times n$ matrix) are orthogonal and unitary, and

$$\Sigma = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \tag{2}$$

where $S = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_r)$, r=min(m, n), each collumn of matrix Uis a orthonormalized eigenvectors of AA^{T}andits counterpart in V the orthogonal eigenvectors of $\mathrm{A}^{\mathrm{T}}\mathrm{A}$, and they are called left singular vector and right singular vector respectively, $\delta_i (i = 1, 2, \ldots, r)$ is called the singular value and is the nonegative square root of eigenvalues of $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ and $\delta_1 \geq \delta_1 \geq \cdots \geq \delta_r \geq 0$. The right singular vectors well reveal the geometry structure of the matrix A, but differ in importance according to the singular values' size, our experiments show that only the first and second singular vector cannot be neglected when constructing the feature vectors. Fig. 4 illustratesthe first and second singular vectors of onemotion matrix ofa motion classwhile Fig. 5 shows first and second singular vectors of all motion matrices of the same motion class(>300).One can easily tell two facts from it:

- Both the first and second singular vectors are insensitive to the motion variations among the same motion class as they fit well with each other.
- The first singular vectors are more insensitive as they vary less than the second singular vectors.

Feature vectors are constructed by concatenating the first and second singular vectors with different weights according to their related singular values.

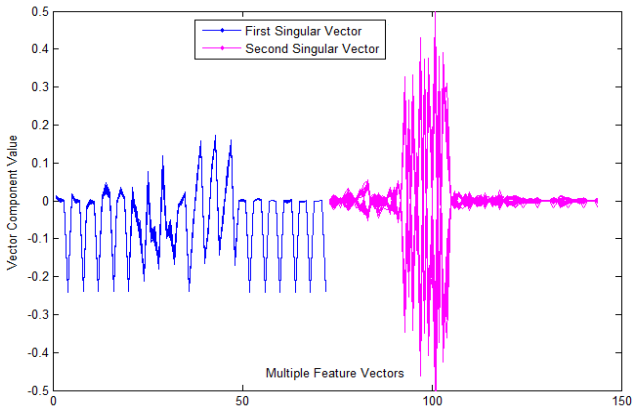

Fig. 5. The first and second singular vectors of all sub-motion in the same sub-motion class.

E. L_1MRecognition by SVM and HMM

SVM has demonstrated many successful applications in pattern recognition problems, which is widely acknowledged the best "off-the-shelf" supervised learning algorithm, where the concept of optimal margin classifier is introduced. Its key idea is finding the optimal hyper plane which maximizes the geometric margin as a decision boundary between two classes of training samples. The geometric margin of the decision boundary with respect to training sample $(x^{(i)}, y^{(i)})$ can be defined as:

$$\gamma^i = y^{(i)}\left(\left(\frac{\omega}{\|\omega\|}\right)^T x^{(i)} + \frac{b}{\|\omega\|}\right) \tag{3}$$

where $(\omega, b)$is the hyperplane, and all the points on it satisfy the equation $\omega^T x + b = 0(\omega = [\theta_1 \ldots \theta_n]^T$and $b = \theta_0$, they are the parameters to be calculated). The problem of maximize the geometric marginis equivalent to the following optimization problem:

$$\min_{\gamma,\omega,b} \frac{1}{2}\|\omega\|^2 \tag{4}$$

$$\text{s.t. } y^{(i)}(\omega^T x^i + b) \geq 1,\ i = 1, \ldots, m$$

The above problem's dual form when introducing Lagrange duality will allow us to use kernels to make the

multi-class classifiable in very high dimensional spaces.

In recognition of L_1M, a one-versus-one multi-class SVM is used to decide which class of the sub-motion belongs to, each classifier is trained using the two classes involved, another method used in multi-class classification with binary classifiers is one-versus-rest, where one class is the positive sample and all the rest are negative samples. As is shown in [13], the former one outperforms the latter one in both accuracy and training time.

Hidden Markov Model is a statistic model which can be built upon a set of unobserved (hidden) states, while the states transitions are related to a Markov process. We use HMM to compute the latent L_1M, which is the hidden state, given the trained parameters of a learning algorithm and a sequence of previous recognized L_1Ms as observations. The set of hidden states in our case is defined as:

$$S = \{\text{state} \mid \text{state} \in L_1\text{M}^{(i)}, \text{i} = 1, \dots, n\} \quad (5)$$

where n is the number of our L_1Ms, and $L_1\text{M}^{(i)}$ is the ith one of all L_1Ms. The set of observation states, in our case, is the same set as S, which is denoted by O. We use a modified version of Viterbi algorithm to implement the HMM and it works well to find the most likely hidden L_1Ms sequence given the observation of the candidate L_1M sequences. The parameters of HMM is trained using Baum-Welch learning algorithm with our real-time recorded training samples.

F. L_2M Recognition by Pattern Matching

One major problem in motion recognition is that logically similar motions are not necessarily numerically similar, there must be spatio-temporal variations between two similar motions executed differently while there also certain aspects of the similar motions which are consistent [14]. Fortunately, we have already recognized a number of L_1Ms, and they are the logically essence of L_2Ms, all we need to do is evaluate the sequence of current recognized L_1Ms and decide which class of L_2M are they in. Table I shows all defined motion patterns.

TABLE I: MOTION PATTERNS OF L_2M

L_2M	*L1Ms involved and their orders*		
Me	Lift left hand	Pat chest	Pat chest
Claps	Clap	Clap	Clap
Wear a Scarf	Lift left hand	Put left arm around neck and back	Put left arm around neck and back
See	Lift left hand	Draw a circle before eyes using left hand	Draw a circle before eyes using left hand
Stop	Stretch out left arm to the left	Palm faces left	N/A
Stretch Oneself	Lift both left and right arms to waist	Stretch out both arms over head	N/A
Left Direction	Lift left hand overhead	Swing left ward and back	Swing left ward and back
Right Direction	Lift right hand over head	Swing right ward and back"	Swing rightward and back

The motion pattern is easily adapted and extended when new motion classes are added.

G. L_3M Recognition by Grammar-Based Command Synthesization

In the synthesization of L_3Ms we introduce the notation for describing the production rules of the motions' grammar, from now, we will use M to denote the generic term of motion, S subject, VI intransitive verb, VT transitive verb, A adverbial adjunct and O object. And we also relate each L_2M to one of the above grammatical roles, for instance, the L_2M (Me) refers to S as subject grammatically, L_2M (Wear a scarf) refers to VI as intransitive verb. The table below is the production rules of our defined grammar.

Together with speech recognized terms, we can synthesize these semantic terms into a sentence. For example, "Me See He Left" which can be interpreted to "I saw him in the left direction", "He Quickly Wears a Scarf" and "He Claps" which are self-explanatory. One can easily think of other examples and the grammar can be extended to accommodate more complex sentences.

TABLE II: GRAMMAR OF MOTION COMMANDS

Rule id	*Production rules*
1	M → S | VI | VT | O | A
2	M → S VI
3	M → S VI A | S A VI
4	M → S VT O
5	M → S VT O A | S A VT O

III. EXPERIMENTAL RESULTS

The method is tested against the bench mark of 8 classes of L_2Ms with their correspondingL$_1$Ms.We conducted the experiments in two ways: one is by using k-fold cross validation where k = 3 and each motion class L_1Mis provided with 300 positive samples and 1000 negative samples, the other one is tested in real-time manner. Four aspects indicating the method's performance has been evaluated: sensitivity, accuracy, specialty and efficiency. While the aspects of different levels are inter-connected, here we only list the results of $L_2Mrecognition$. Table III shows the experimental results of both k-fold cross validation and real-time testing, where CV denotes cross validation and RT denotes real-time. Table IV shows the comparison with other methods in accuracy.

Experiments show that the proposed method has achieved remarkable performance with the average sensitivity reaching 90.6%, average specialty 95.7% and accuracy 95.4%;Our approach can handle motions with similar sub-motions, for instance, L_2M motion "Me", "Wear a scarf" and "See" all contain the sub-motion of "Lift left hand" which is L_1M, the similar sub-motions can offer probability information of the three motions if handled properly, otherwise, it will be mixed up and result in false recognition.

The test platform's software/hardware configuration is as follows: 64-bit Windows 7 Enterprise, Intel® Core™2 Duo CPU E7500 @2.93GHz (Duo) and 4GB RAM. The average time consumption of each L_2Mis 6ms, much faster than the real-time motion data providing, which is at a rate of 50 fps (20ms each frame), thus the efficiency of our approach is

guaranteed.

TABLE III: EXPERIMENTAL RESULTS

L_2M	*Sensitivity(%)*		*Specialty (%)*		*Accuracy (%)*	
	CV	RT	CV	RT	CV	RTs
Me	88.4	76.1	98.7	92.4	92.8	91.3
Claps	97.6	89.4	96.6	94.7	98.4	97.6
Wear a Scarf	100	95.8	100	100	100	97.8
See	95.5	86.7	100	97.2	100	96.8
Stop	100	93.5	97.8	94.3	100	94.5
Stretch Oneself	100	91	100	93.4	100	92.3
Left Direction	100	100	100	100	100	100
Right Direction	100	92.6	100	93.5	100	93.1

TABLE IV: COMPARISON WITH OTHER METHODS

Method	Accuracy
Vieira *et al.* [15]	84.8%
Raptis *et al.* (only correlation) [5]	89.9%
Raptis *et al.* (energy + correlation) [5]	96.9%
Our approach	95.4%

IV. CONCLUSION

In this paper, we proposed a hierarchical method to recognize three different levels of motions and obtained good performance.SVD together with SVM show great power of multi-class classification, HMM and pattern matching are flexible tools to capture and cope with spatio-temporal variations of similar motions. In a high level of motion recognition, grammar-based method can be an appropriate choice. Future works will be focused on finding other good features in motions as well as facilitating training data collection.

ACKNOWLEDGMENT

We would like to thank members of DALab in School of Software who offer valuable advices and help in implementing part of the recognition framework and data collection.

REFERENCES

[1] R. M. Baecker and W. Buxton, *Human-Computer Interaction: A Multidisciplinary Approach*, 1987.
[2] C. Li, P. R. Kulkarni, and B. Prabhakaran, "Segmentation and recognition of motion capture data stream by classification," *Multimedia Tools and Applications*, vol. 35, no. 1, pp. 55-70, 2007.
[3] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Trans. on Graphics (TOG)*, pp. 559-568, 2004.
[4] J. Alon, V. Athitsos, and S. Sclaroff, "Accurate and efficient gesture spotting via pruning and subgesture reasoning," *Computer Vision in Human-Computer Interaction*, pp. 189-198, 2005.
[5] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2011, pp. 147-156.
[6] H. K. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1999, vol. 21, no. 10, pp. 961-973.
[7] D. Kim, J. Song, and D. Kim, "Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs," *Pattern Recognition*, vol. 40, no. 11, pp. 3012-3026, 2007.
[8] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. 2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 444-451.
[9] Q. Gu, J. Peng, and Z. Deng, "Compression of human motion capture data using motion pattern indexing," in *Proc. Computer Graphics Forum*, 2009, pp. 1-12.
[10] M. Ryoo and J. Aggarwal, "Stochastic representation and recognition of high-level group activities," *International Journal of Computer Vision*, vol. 93, no. 2, pp. 183-200, 2011.
[11] A. F. Bobick, "Movement, activity and action: The role of knowledge in the perception of motion," *Philosophical Trans. of the Royal Society of London. Series B: Biological Sciences*, vol. 352, no. 1358, pp. 1257-1265, 1997.
[12] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, 1970, vol. 14, no. 5, pp. 403-420.
[13] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. on Neural Networks*, 2002, vol. 13, no. 2, pp. 415-425.
[14] M. Muller and T. Roder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proc. 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Vienna, Austria, 2006, pp. 137-146.
[15] A. W. Vieira *et al.*, "Stop: Space-time occupancy patterns for 3dactionrecognition from depth map sequences," *CIARP*, 2012.

Jianchao Lv was born in Shandong Province on March 30, 1989. He received BS degree in software engineering from Software Engineering Institute of East China Normal University, Shanghai, China. His research interests include human computer interaction and machine learning. He is now a second year graduate student in School of Software, Shanghai Jiaotong University pursing his Master Degree of Engineering.

Shuangjiu Xiao was born in Sichuan Province on September 9, 1973. She received Ph.D. degree of Computer Aided Design in 2002 from Northwestern Polytechnical University, Xi'an, China. The major field of study of her includes computer graphics, human computer interaction.

She is an associate professor of School of Software of Shanghai Jiao Tong University. She did Postdoctoral research during 2002-2004 in Computer Science and Technology Department of SJTU. She has about 70 papers published.

Prof. Xiao is a member of the China Graphics Society and the China Computer Federation.