Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

A Novel Method to Recognize Complex Dynamic Gesture by Combining HMM and FNN Models

Xiying Wang[1] Guozhong Dai[1]
[1]*Institute of Software, Chinese Academy of Sciences, Beijing 100080, China*
E-MAIL:wangxy0823@sohu.com or xiying04@ios.cn,

Abstract: *Recognition of dynamic gesture is an important task for gesture-based Human-Computer Interaction. A novel HMM-FNN model is proposed in this paper for the modeling and recognition of complex dynamic gesture. It combines temporal modeling capability of Hidden Markov Model, and ability of Fuzzy Neural Network for fuzzy rule modeling and fuzzy inference. Complex dynamic gesture has two important properties: its motion can be decomposed and usually being defined in a fuzzy way. By HMM-FNN model, dynamic gesture is firstly decomposed into three independent parts: posture changing, 2D motion trajectory and movement in Z-axis direction, and each of part is modeled by a group of HMM models which represent all fuzzy classes it possibly belongs to. The likelihood probability of HMM model to observation sequence is considered as fuzzy membership for FNN model. In our method, high dimensional gesture feature is transformed into several low dimensional features, which leads to the reduction of model complexity. By means of fuzzy inference, it achieves a higher recognition rate than conventional HMM model. Besides, human's experience can be taken advantaged to build and optimize model structure. Experiments show that the proposed approach is an effective method for the modeling and recognition of complex dynamic gesture.*

1. Introduction

Visual gesture has the potential to be a natural and powerful tool in many applications, supporting efficient and intuitive interaction between human and computer. Gesture recognition has becomes a very important research direction in the filed of computer vision and Human-Computer Interaction [1~4]. Although gesture can be categorized in several different ways [1], when considered from the view of motion feature, gesture usually could be categorized into static and dynamic gesture. Static gesture, also called posture, is defined as a static movement, such as "okay" sign. It expresses meaning just by means of hand shape or finger configuration, and thus it can be thought as special case of dynamic gesture. Dynamic gesture is defined as a dynamic movement, and it can further classified into simple and complex gesture. Simple gesture involves a fixed posture and change in the position or orientation of the hand, such as making a pinching posture and changing the hand's position. Complex dynamic gesture is the one includes not only the movement of hand but also the change of fingers.

Unlike most of previous works that mainly focus on recognition of static gesture (posture) or simple dynamic gesture, we will focus our interests on complex dynamic gesture recognition in this paper. Generally speaking, dynamic gesture owns two characteristics as follows: (1) temporal variability. The speed of gesture performing is different from person to person. (2) Spatial variability. The position of gesture and its movement range are variable. Furthermore, for complex dynamic gesture, we find there are still two additional characteristics of them: (1) Complex dynamic gesture is decomposable structurally, in other words, they can be decomposed into three parts: the change of posture or fingers' configuration, 2D movement of whole hand and movement component in the Z-axis direction. (2) Description or definition of dynamic gesture always involve some indefinite linguistic factors. In most cases, people used to define a gesture by some ambiguous words instead of strict definition, such as the following example: a dynamic process for hand posture changing from all fingers extended to a fist, meanwhile the hand trajectory is like a circle, is defined as the gesture of "all selected". The fuzzy characteristic of dynamic gesture motivates us to apply fuzzy set theory to gesture recognition.

In this paper, we propose a novel approach for the modeling and recognition of complex gesture based on HMM-FNN model. The proposed method takes full advantage of the characteristics of complex dynamic gesture, i.e. it can be decomposed and often defined by ambiguous words. HMM-FNN model combines the advantage of HMM, that is, the strong modeling ability for temporal data, and capability of FNN model for fuzzy rule modeling and fuzzy inference. By HMM-FNN model, the complex gesture is decomposed firstly into three independent parts, each of which is modeled by a series of HMM models, and then get the final recognition result by fuzzy inference of FNN. By the proposed method, the high-dimensional feature of complex dynamic gesture is reduced to three low-dimensional features, with the consequence of reduction of computational cost, and human's prior knowledge about gesture definition can be used to help construction and optimization of model structure.

The organization of this paper is as follows. Second section introduces related work for methods of gesture recognition. The third section describes our method of gesture feature extraction in the condition of monocular vision. The fourth section introduces the proposed HMM-FNN model and its training process. In fifth section, experiments are performed to verify the effect of proposed method, and the comparison result between our method and conventional method is listed. The conclusion and future work are given in the last section.

2. Related Works

In 1990's, Starner[5] applied HMM method to the recognition of American Sign Language. In his method, there are four features, i.e. hand's x and y position, angle of axis of least inertia and eccentricity of bounding ellipse, used to form the feature vector of simple gesture. Rigoll etc. [6] compute the difference image between frames and construct a 7-dimensional vector as the representation of dynamic gesture. Since the feature is merely from the difference image, it's difficult to get full information about hand gesture by such method. Bobick and Wilson [7] used a state-based method to model and recognize hand gesture, and described gesture as a trajectory in the feature space of hand. By dividing the trajectory into several segments, the gesture could be modeled as a sequence of states. Different from HMM based method, in which the parameters of model is trained by samples, Bobick's method models the gesture by an example and achieves recognition by means of matching. Ren and Zhu [8,9] built gesture's appearance model by combining color information and motion information of the gesture, and applied Dynamic Time Wrapping (DTW) to the recognition of gesture. HyeSun Park etc. [10] proposed a method to build feature vector by the relationship between hand region and face, and a integrated HMM model was applied to recognize hand gesture instead of a series of single HMM models. The main drawback of this method is the complex model structure, which would lead to a low-efficient training and recognition process. Besides, most previous methods of gesture recognition only take advantage of basic image information of gesture video, and high-level cues, such as human experience or rules, are ignored.

Neural network is usually applied to the recognition of static hand gesture [11,12]. There are also researchers [13] who have tried to apply Neural Network to the dynamic gesture recognition, but the capability of Neural Network on modeling of temporal data is very limited and often leads to high computation cost, thus it's difficult for Neural Network to become popular method of dynamic gesture recognition. However, since Neural Network has stronger ability for parameter estimation than common statistical methods, it has been used to estimate the posterior probability of HMM model. Morgan and Bourlard [14] begin such research by combining Multi-layer Perceptron with HMM model. Michael Cohen and his colleagues [15] also suggest that the combination of HMM and Neural Network can relax the limitation of the assumption that features of observation vector are independent one another.

3. Feature Extraction of Dynamic Gesture

The movement of 3D dynamic gesture can be decomposed into three parts: sequence of changing postures, 2D motion trajectory and movement in the Z-axis direction. With the purpose of simplification, we describe the movement component in the Z-axis direction by the changing of palm area S_p, in other words, when S_p value becomes larger and larger, it implies that hand is moving towards camera, or else, hand is moving away from camera.

The feature of posture we used is a one-dimensional value, which is obtained based on our previous works about hand and fingertip tracking. In our previous works, we have accomplished hand segmentation from background [17], and the continuous tracking of deformable hand gesture [18]. To deal with the continuously changing hand shape, CamShift algorithm is applied to whole hand tracking and Particle Filter for multi-finger tracking. At the same time, tracking detection process can update tracking template dynamically in order to assure a successful continuous tracking. Fig.1 shows the effect of our tracking method, in which the larger circle indicates the position of palm and the two little circles show the position of two fingertips.

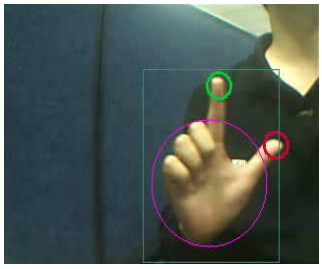

Fig.1 Gesture tracking

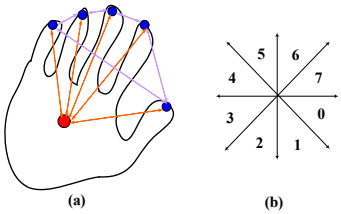

Fig.2. Calculation of gesture feature

The feature of posture is calculated by the length of all visible fingers and the distances among all adjacent fingers as shown in Fig.2(a). By adding a different value V_{base} for different posture type, feature value of different types of posture can be spread in different ranges. Eq.1 gives the calculation equation of posture feature value.

$$V = \alpha[V_{base} + 100 * \sum_{i=0}^{4}(Len(i)/D_h) + 100 * \sum_{j=0}^{n}(Dis(j)/D_h)] + (100-\alpha)E \quad (1)$$

where D_k is the diameter of bounding circle of hand region in order to normalize the length and distance value, n is the number of visible fingers and E is the eccentricity of hand region, whose calculation is shown by Eq. 2.

$$E = \sqrt{\frac{(A+B) - \sqrt{(A-B)^2 + 4H^2}}{(A+B) + \sqrt{(A-B)^2 + 4H^2}}} \quad (2)$$

$$A = \sum m_i(y_i^2 + z_i^2), B = \sum m_i(z_i^2 + x_i^2), H = \sum m_i x_i y_i$$

where A and B are moments of inertia around x and y axis, H is the inertia product. The value of E is invariant to transition, rotation and scaling.

The feature of 2D motion trajectory of hand gesture is represented by a series of discrete movement direction value. For the 2D motion plane, we divide the direction into 8 discrete values as shown in Fig.2(b). Thus, the trajectory of dynamic gesture can be described by the sequence of discrete direction value R: $r_1, r_2, ...r_i.....r_T$, $(0 \le r_i \le 7)$.

The feature of gesture movement component in the Z-axis direction is represented by the changing of palm area. The value of palm area should be normalized by comparison with the palm area of first frame in the gesture video. As a result, movement in the Z-axis direction is described by the sequence S: $s_0, s_1, ...s_i$, $(0 < s \le 1.0)$.

4. HMM-FNN framework for Gesture modeling and recognition

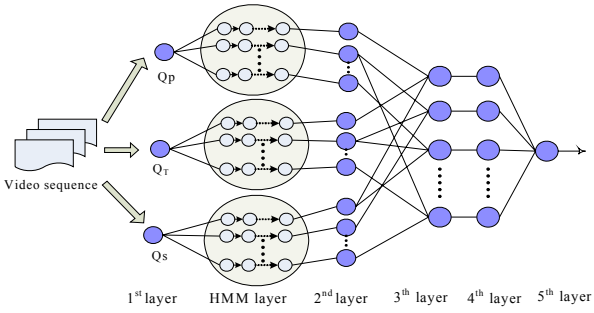

Fig.3 HMM-FNN model

4.1 HMM-FNN model

HMM-FNN model, as shown by Fig.3, is the combination of Hidden Markov Model and Fuzzy Neural Network. It's well known that HMM model has strong ability for temporal data modeling, so we apply it to the modeling of the three movement components of complex dynamic gesture. Fuzzy Neural Network has strong ability for fuzzy rule modeling and fuzzy inference due to its integration of fuzzy set theory and Neural Network together. Since traditional FNN cannot model temporal data and conventional HMM do not own ability for fuzzy inference, we integrate the two models together to represent complex dynamic gesture and perform inference by the integrated HMM-FNN model, which is shown in Fig.3, for the recognition of dynamic gesture.

HMM-FNN model includes five layers. Its first layer, second layer and HMM layer constitute the fuzzy preprocessing part, third layer and fourth layer constitute fuzzy inference part, fifth layer is the defuzzification part of HMM-FNN and produce distinct output. The following will introduce these five layers in detail.

The first layer is the input layer of the model and it has three neurons, which correspond to the three movement components of dynamic gesture, i.e. Q_P, Q_T and Q_S, respectively.

The second layer and HMM layer compose fuzzification layer. Each HMM model is related to a neuron in second layer, which represent a fuzzy class to which the input observation possibly belongs. The likelihood of input observation sequence Q to each HMM model, i.e. $p(Q/\lambda)$, is considered as membership value of the corresponding fuzzy class variable. At the same time, the neurons in second layer constitute the antecedent part (conditional part) of fuzzy rule. The number of neurons of this layer is $m_1+m_2+m_3$, where m_1, m_2, m_3 are the class numbers of posture changing, 2D trajectory and movement in the Z-axis direction respectively.

The third layer is the layer of fuzzy inference, and each neuron represents a fuzzy rule. An example of fuzzy rule maybe like this one: If posture changing is similar as the process that all extended fingers deflex to a fist (condition C1), the trajectory of hand is like a circle (condition C2) and the palm area almost keep constant (condition C3), then the type of dynamic gesture is A (rule conclusion), i.e. $C_1 \wedge C_2 \wedge C_3 \rightarrow A$. The connecting weights between neurons in second and third layer imply the contribution degree of the antecedent part for this rule. The output of neuron in third layer is calculated as shown in Eq.3.

$$O^{(3)} = b = \sum_{i=0}^{m} \omega_i I_i^{(3)} = \sum_{i=0}^{m} \omega_i p(Q/\lambda_i), \quad and \sum \omega_i = 1 \tag{3}$$

The fourth layer is normalization layer, the neuron number of which is equal to that of third layer. In order to speed up convergence of the network during training, the output of third layer is normalized to assure the sum of them is equal to 1. Output of its neuron is shown as Eq.4.

$$O_i^{(4)} = I_i^{(4)} / \sum_{i=0}^{N} I_i^{(4)} \tag{4}$$

The fifth layer is the defuzzification layer, the output of which is shown as Eq.5.

$$O^{(5)} = \sum_{j=1}^{N} \omega_j O_j^{(4)}, \ and \ \sum_{j=1}^{N} \omega_j = 1 \tag{5}$$

where ω_j implies the importance of each rule for the final classification output, N is the total number of fuzzy rules.

We choose left-right Bakis model [16] as the type of HMM model due to its straightforward structure. Corresponding to the features' type, the type of HMM models for posture changing and movement in Z-axis direction are one-dimensional continuous HMM models, while that of 2D trajectory is a one-dimensional discrete one. As for continuous HMM model, we employ Gaussian Mixture Model (GMM) as the emission probability of observation, which has the likelihood as described in Equ.6:

$$p(O/\lambda) = \sum_{i=1}^{M} \omega_i g_i(x) \tag{6}$$

where ω_i is the weigh of i^{th} Gaussian component.

4.2. Training of gesture model

Training of HMM-FNN model includes two phases: Firstly, the training of HMM model, that is the re-estimation of parameters in state transition matrix, output probability

matrix or GMM's expectation and variance. We apply Baum-Welch method [16], which is based on Expectation-Maximum (EM) algorithm, to perform training of HMM models, and obtain the final parameters according to the Maximum Likelihood (ML) criteria.

Secondly, the weights of FNN will be trained after HMM training. It involves the weights used in Eq.3 and in Eq.5. Back-Propagation (BP) algorithm is chosen for the training of connecting weights. According to BP algorithm, we can get the formula for weight adjustment as follows:

$$\Delta\omega_k^{(5)} = -\eta\delta^{(5)} \bullet O^{(5)} = -\eta(t-z)I_k^{(5)}$$
$$\Delta\omega_{ij}^{(3)} = -\eta I_i^{(3)}\delta_j^{(3)} = -\eta(\omega_j^{(5)}\delta^{(5)})I_i^{(3)} = -\eta(\omega_j^{(5)} \bullet (t-z))I_i^{(3)} \quad (7)$$

Updated weight value can be calculated by $w(T+1) = w(T) + \Delta w$. When it reaches maximum iteration number or converge, the training process stop.

4.3. Gesture recognition

Suppose that complex gesture has already been decomposed into three independent parts during hand tracking. The three feature sequences are considered as input of HMM-FNN model, and calculate the likelihood of HMM model $p(O/\lambda)$ according to forward probability method [16]. The final output of HMM-FNN model indicates the class type to which the input gesture belongs, such as the output of gesture A is between the range $(0,2]$, and gesture B is between $(2,4]$ and so on.

5. Experiments

Fig.5 and the other four are upwards, downwards, leftwards and rightwards trajectories. Besides, there are four types of movement in the Z-axis direction (Q_S,) defined in our system: moving towards camera, moving away from camera, moving towards and then away camera, and keeping constant.

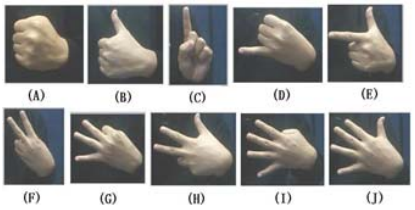

Fig.4 Ten pre-defined postures

Based on the definition of all classes of Q_P, Q_T and Q_S, we design 14 different complex dynamic gestures for the task of interaction between human and computer. Fig.6 gives three of them by extracting several frames among the sequence of gesture images. In each frame of Fig.6, hand and fingertip's positions, which are obtained by tracking, are demonstrated by colored circles.

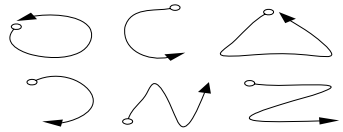

Fig.5 Six pre-defined trajectories

The gesture shown by Fig.6(a) is the one that posture changes from Posture-E to Posture-B, and at the same time hand is moving towards camera. The gesture shown by Fig.6(b) is the one that posture changes from fist to Posture-B and then to Posture-E, and at the same time hand moves rightwards. The gesture shown by Fig.6(c) is the one that posture changes from fist to Posture-B and then back to fist, at the same time hand moves along a circle-like trajectory.

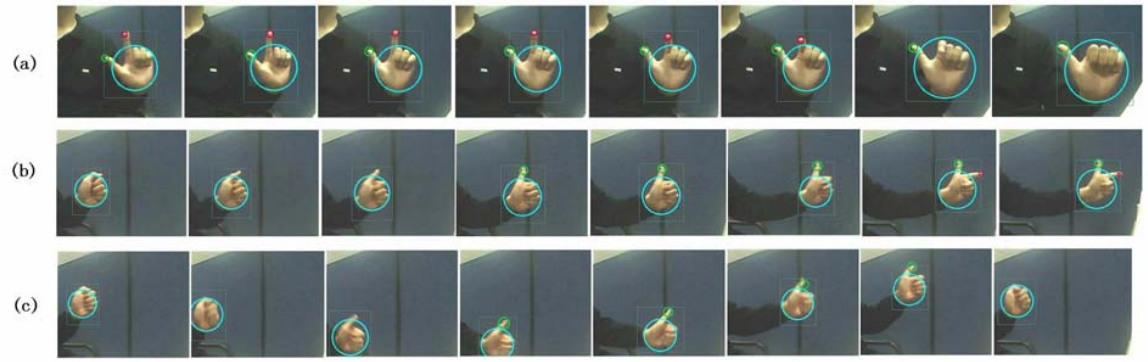

Fig.6 Three examples of complex dynamic gestures

We implement the proposed method by C++ language, and do a series of experiments on the common PC with CPU of Pentium IV 1.7G and 256M memory. Firstly, we define ten basic postures demonstrated by Fig.4, and marked by Posture A~J respectively. There are totally 8 kinds of posture changing process (Q_P) defined in our experiments, such as the process of posture changing from Posture-E to Posture-A and so on. That is, there are eight kinds of HMM models in HMM-FNN which are responding to posture changing.

As for the trajectory of 2D hand movement (Q_T), ten types have been defined, six of which have been demonstrated by

By Eq.1, we calculate the feature values of above three gestures. As shown in Fig.7(a), the blue curve, orange curve and yellow curve indicate the changing posture feature values of dynamic gestures demonstrated in Fig.6(a), (b) and (c) respectively. In Fig.7(b), the three curves indicate the sequences of palm area of the three dynamic gestures respectively.

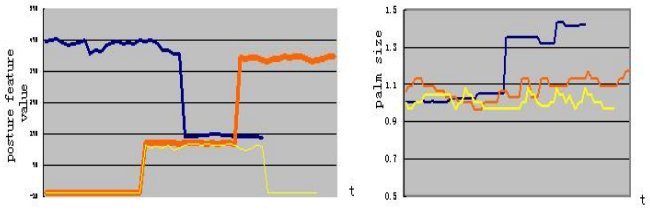

Fig.7 (a) Curves of posture feature value. (b) Curves of palm area

In our experiments, the fuzzy rules, with a total number of 24, are obtained by data clustering combined with human experiences. The initial connecting weights are also set by people's prior knowledge about dynamic gestures. For each dynamic gesture, we ask each of five users to perform it 3 times, and then get 15 video clips. As a result, there are totally 210 training samples for all 14 gestures, 180 of which are used for model training and the others are for testing. When the error of testing is below the threshold or training times reach its maximum, the model is considered well-trained. We evaluate the performance of proposed model by comparison with conventional HMM model, and get the results as shown in Table.1.

Table.1 Comparison of recognition rate between proposed method and conventional HMM model.

Gesture type	A	B	C	D	E	F	G
Conventional HMM	73.3 %	70.0 %	73.3 %	80.0 %	70.0 %	83.3 %	70.0 %
Proposed HMM-FNN	86.6 %	86.6 %	83.3 %	90.0 %	86.6 %	90.0 %	80.0 %
Gesture type	H	I	J	K	L	M	N
Conventional HMM	66.6 %	80.0 %	73.3 %	70.0 %	70.0 %	76.6 %	70.0 %
Proposed HMM-FNN	83.3 %	80.0 %	83.3 %	90.0 %	83.3 %	83.3 %	86.6 %

From the experiment results, we can see that the proposed model has a better recognition performance than conventional HMM model, and can achieve a satisfying result for recognition of complex dynamic gesture. Furthermore, experiments show the recognition rate of those HMMs embedded in HMM-FNN model for posture changing, 2D trajectory and movement in the Z-axis direction are 76.6%、80.0% and 73.3% respectively, but the final recognition rate of HMM-FNN is 88.3%. Thus, it can be concluded that the recognition error of HMM model can be corrected by fuzzy inference of HMM-FNN model to some extent.

6. Conclusion and Future Work

In this paper, we propose a method based on HMM-FNN method for the modeling and recognition of complex dynamic gesture. It takes full advantage of characteristics of complex gesture. Firstly, it decomposes complex gesture into three independent parts. Secondly, it models complex gesture by fuzzy rules and recognize them through fuzzy inference. The proposed model, HMM-FNN model, combines the capability of HMM model for temporal date modeling, and advantages of FNN for fuzzy rule modeling and fuzzy inference, and even the self-learning ability of Neural Network.

Compared with conventional HMM model, our method reduces the computational cost by decomposing one high-dimensional feature into three low-dimensional temporal features. Besides, through fuzzy inference, the robustness of recognition system is enhanced and achieves a higher recognition rate. What's more, human experience can be employed in the process of model building and optimization.

In our future work, we are planning to use temporal and spatial context information of gesture in order to achieve better segmentation and interpretation of complex dynamic gesture.

Acknowledgement

This work was supported by National Basic Research Program (973 Program) of China under Grant No. 2002CB312103, and 863 Program under Grant No. 2006AA01Z328.

References

1. Valdimir I. Pavlovic, R. Sharma, T. S. Huang. Visual interpretation of hand gesture for Human-Computer Interaction: a review. IEEE PAMI, vol.19(7), 1997. 677~695.
2. Ying Wu and Thomas S. Huang. Vision-based gesture recognition: a review. Gesture workshop, 1999. 105~115.
3. Guangqi Ye, Jason J. Corso, Gregory D. Hager. Gesture recognition using 3D appearance and motion feature. In proceedings of IEEE CVPR'2004. 160~170.
4. Micah Alpern and Katie Minardo. Developing a car gesture interface for use as a secondary task. CHI'2003. pp.932~933.
5. Thad Starner and Alex Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. Technical Report 375, MIT media Lab, Perceptual Computing Group. 1995.
6. Gerhard Rigoll, Andreas Kosmala and Stefan Eickeler. High performance real-time gesture recognition using Hidden Markov Models. Gesture Workshop. 1997. 69~80.
7. Aaron F. Bobick and Andrew D. Willson. A state-based approach to the representation and recognition of gesture. IEEE Trans. on Pattern Analysis and Machine Intelligence. 1997. 19(12):1325~1337.
8. REN Hai-Bing, ZHU Yuan-Xin, XU Guang-You, LIN Xue-Yin and ZHANG Xiao-Ping. Spatio-temporal appearance modeling and recognition of continous dynamic hand gestures (in Chinese with English abstract). Chinese Journal of Computers. 2000. 23(8):824~828.
9. Yuanxin Zhu, Guangyou Xu and David J. Kriegman. A real-time approach to the spotting, representation, and recognition of hand gesture for Human-Computer Interaction. Computer Vision and Image Understanding. Vol.85, 2002. pp.189~208.
10. HyeSun Park, EunYi Kim, SangSu Jang and HangJoon Kim. An HMM-based gesture recognition for Perceptual User Inerface. The fifth Pacific-Rim Conference on Multimedia(PCM 2004). LNCS 3332, 1027~1034.
11. Kouichi Murakami and Hitomi Taguchi. Gesture recognition using recurrent neural network. In proceedings of CHI'1991. 1991, 237~241.
12. Sidney Fels and Geoffrey Hinton. Glove-TalkII: an adaptive

gesture-to-formant interface. In proceedings of CHI' 1995. 456-463.

13. Peter Vamplew and Anthony Adams. Recognition and anticipation of hand motions using a recurrent neural network.. In proceedings of IEEE International Conference on Neural Networks. 1995. 2904~2907.
14. N. Morgan and H. Bourlard. Continuous speech recognition using multilayer perceptrons with Hidden Markov Models. In proceedings of ICASSP'1990. 413~416.
15. Michael Cohen, David Rumelhart, and etc. Combining Neural Networks and Hidden Markov Models for continuous speech recognition. In ARPA Continuous Speech Recognition Workshop. Stanford University, CA. Sep, 1992.
16. Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of IEEE. 1989, 77(2): 257~286.
17. Jiyu Zhu, Xiying Wang, Weixin Wang and Guozhong Dai. Hand gesture recognition based on structure analysis. Accepted by ChinaGraph'2006.
18. Xiying Wang, Xiwen Zhang and Guozhong Dai. A novel approach to tracking deformable hand gesture for real-time interaction. Accepted by Chinese Journal of software.