Workshop on Multimedia Technology

Hand gesture mining

Xu Yan[1] Huai Jin-Peng[1] Wang Zhaoqi[2]

1. Department of computer Science and Engineering,
Beijing University of Aeronautics and Astronautics, Beijing 100083,China;
2. Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100080, China;

Email: xu.yan@263.net, Huai@cscw.buaa.edu.cn, zqwang@ict.ac.cn
Tel: +86-10-82681762, +86-10- 82316262, +86-10-82649121

Abstract

Hand gesture recognition and synthesis are both important for more natural and immersive Human Computer Interaction, many approaches to these tasks involve a large hand motion search space. Human hand motion is highly articulate and highly constrained. Although some constrains can be obtained by hand anatomy and biomechanics, it is an approximation of natural finger motion, in addition, these exist many constrains, which cannot be explicitly represented. In this paper, we presents a data mining approach to discovery the constraints of human hand motion, the result of the data mining is presented to display by VRML. Our presented approach is helpful for analyzing hand motion and hand motion capturing.

Keyword: Human Computer Interaction, Hand gesture recognition, Data mining, Rough Set

1. Introduction

In recent years, hand gesture recognition is one of the most important parts of gesture interface. In many virtual environment applications, hand gesture plays an important role in gesture recognition and related work due to interest in a more natural and immersive Human Computer Interaction (HCI).

In another way, The synthesis of a sign language is to generate the sign animation from a text, which can be used to assist the deaf people to watch TV, browse Internet, study, and help the deaf people communicate with the hearing people conveniently. The Census by government of China in 1987 enumerated that 2.18 % of the total population were deaf. It predicted that there would be about 28 million of deaf people in China up to now. Most of them communicate with each other by Chinese Sign Language. We would like to fill in the gaps between the deaf and hearing using multi-modal interface technology in a project.

Hand gesture recognition and synthesis of sigh language are both important for more natural and immersive Human Computer Interaction, many approaches to which involve a large hand motion search space. However, a comprehensive review of various techniques in hand modeling, constrains finding is needed. Human hand motion is highly articulate, because the hand consists of many connected parts leading to complex kinematics. At the same time, hand motion is also highly constrained, which makes it difficult to model[1].

Figure 1 shows the skeleton of a hand. Each finger consists of tree joints. Except for the thumb, there are 2 degrees of freedom(DOF) for metacapophalangeal(MCP)joins, and 1 DOF for proximal interphalangeal(PIP)joins and distal interphalangeal(DIP) joints. We can assume 2 DOF for the MCP and TM joint, and 1 DOF for all the other joints. Generally, the hand roughly has 21 DOF.

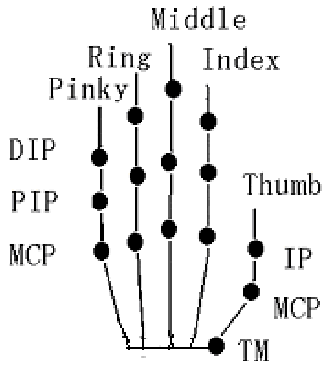


Figure 1 Kinematical structure and joint notions

A Hand has 21 DOF, and Hand motion make a large state space. However, although the human hand is a highly articulated object, it is highly constrained.

Many researches focus on the Constraints of Human hand motion. One type of constrains is the limits of the range of finger motions as a result of hand anatomy, $0^o \le \theta_{MCP} \le 90^o$, $0^o \le \theta_{PIP} \le 110^o$ .The motion of DIP joint and PIP joint follows such constrains as $\theta_{DIP} = (2/3)\theta_{PIP}$ from the study of biomechanics [2][3], these constrains were intentionally made invalid, it is an approximation of nature finger motion.

A useful approach to this task was proposed by John Lin, Ying Wu and Thomas S.Huang, they proposed a learning approach, finger motion observed in the real motion data collected by CyberGlove[4]. One of the common problems of the learning approaches is that learning result largely depends on the training data set.

In this paper, we use hand motion data collected by CyberGlove, and propose a data mining approach to cluster hand gesture and then use rough set approach to reduce the possible hand space. The main advantage of this approach is that the constrains can be directly made by nature finger motion. In addition, we have collected enormous sums of hand motion data.

2. Data preparing

Hand gesture can be classified into several categories according to different application,. Sign language is an important hand gesture, according to some papers sigh languages are highly structured [5][6], and they are suitable as a test-bed. In this paper, we describe hand motion by Chinese Sign Language (CSL), the hand real motion data collected by CyberGlove.

1.1 Representation of the motion of CSL

In this section, we will introduce the representation of the motion of Chinese Sign Language. We used a joint-based hand model to represent a motion of the hands, with the exception of the shoulder, elbow and wrist, each palm has 21 DOFs. In other words, a vector with 21 elements can represent the gesture of any hand, and the motion of any hand can be represented as follows:

$$\mathbf{G}(t) : \mathbf{R} \rightarrow \mathbf{R}^{21}$$

And a single DOF of a hand can be represented by a function, Gi (t), which is continuous because of the coherence of hand motion. So, each DOF can be represented by a continuous curve as shown in Figure 2.

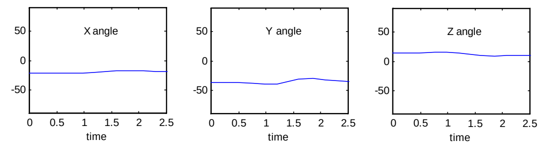


Figure 2. 3 DOFs for thumb'TM motion of sign word "all" (Chinese)

2.2 Animating Chinese Sign Language

To generate the motion of a sentence, we should concatenate the motion-snippets of every sign words. To display this generated motion, we need to interpolate frame between two continuous frames. These two frames may be in the same motion-snippet of one sign word, or the first frame is the end frame of a sign word and the second frame is the first frame of another sign word.

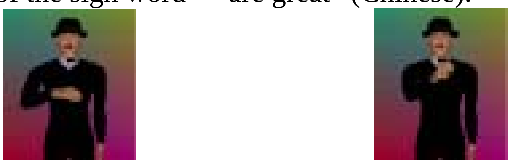
In the first case, linear interpolation can get satisfactory results because of the coherence between these two continuous frames. Let f1 and f2 be two continuous frames of a sign word, tf1 and tf2 be the time of these two frames respectively. For a DOF curve Gi(t) , its value on time tf' $\in$ (tf1 , tf2)can be calculated by the following linear interpolation:

$$G_i(t_{f'}) = G_i(t_1) + \frac{t_{f'} - t_{f_1}}{t_{f_2} - t_{f_1}}(G_i(t_{f_2}) - G_i(t_{f_1}))$$

In the second case, it is more difficult to interpolate when two frames come from two different sign words because these two frames have no coherence. We use an interpolation method based on quaternion[8]. Let f1 and f2 be two frames of the motion a sign sentence, tf1 and tf2 be the time of these two frames respectively. If a joint's orientations on these two time are represented by two quaternions, qf1 and qf2, respectively, then its orientation on time tf' ∈ (tf1 , tf2] can be calculated by the following interpolation:

$$q_{t_{f'}} = \frac{\sin(1 - t_{f'})\theta}{\sin\theta} q_{f_1} + \frac{\sin(t_{f'}\theta)}{\sin\theta} q_{f_2}$$

Where θ is determined by equation $q_{f_1} \cdot q_{f_2} = \cos\theta$. Figure 3 shows the interpolation results between the end frame of the sign word "all of you" (Chinese) and the first frame of the sign word "are great" (Chinese).



(a) The first frame of "all of you" and the last frame of "are great"

(b) Interpolated frames between these two sign words.

Figure 3. Interpolation between two different sign words.

2.3 Building the motion database of Chinese sign words

There are two methods of creating a motion database of sign words, of using an animation software package and customizing it for fast sign transcription, and of using motion capture which records the motion of hands through data gloves and sensors. Each has its own advantages and disadvantages. The method of motion capture can get more realistic hand motion rapidly for sign language synthesis but most recorded gesture are not good enough to use directly, and should be modified, for example, by cutting or tuning fine. On the other hand, if we create a database interactively by using an animation software package, we can get hand motion more exactly but at the same time put a burden of work on the users.

We used a mixed method of creating the database of Chinese sign words. At first, we record the motion of 5500 sign words of CSL selected in the handbook of Chinese Sign Words, and then edit the motion of every word we recorded by using a motion edition software, GestureEdit, which we developed. GestureEdit can cut any frame of a motion or refine the handshape and position of both hands in a frame. Of course, when modifying a motion of a sign word, we need only modify some key frames manually, not all frames. These unmodified frames can be modified automatically by the system based on the coherence between 2 continuous frames. We call these modified frames control points, and call this method a control-points-based motion editing method.

3. Visualization based on VRML

The Virtual Reality Modeling Language (VRML) is a language for describing multi-participant interactive simulations[11] -- virtual worlds networked via the global Internet and hyper-linked with the World Wide Web.

To animate a sign word or a sentence, we use a VRML-based human model, which can be downloaded from Internet or created in a general animating software such as 3Dmax, Softimage, etc., to display each pose of a virtual human.

To represent a virtual human and support interaction between virtual human created by different organizations, a specific sub-language, H-Anim[12]is proposed, which has the same syntax with VRML, and is specific to represent a virtual human.

There are 3 types of nodes in H-Anim to represent a human body: one humanoid, 77 joints, and 47 segments, as shown in figure 4. Joints defined the posture, and segments defined the geometrical model of every segment such as arms, legs and fingers.

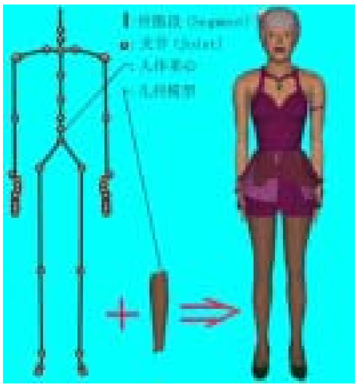


Figure 4 virtual human representations in VRML

In our system, all joints represented by an array with 96 elements, each of them defined a DOF of these joints:

1) AngleValue[0]~ AngleValue[31] : 32 DOFs of left hand.

2) AngleValue[32]~ AngleValue[63]: 32 DOFs of right hand.

3) AngleValue[64]~ AngleValue[69]: 6 DOFs of left leg.

4) AngleValue[70]~ AngleValue[75]: 6 DOFs of right leg.

5) AngleValue[76]~ AngleValue[81]: 6 DOFs of head

6) AngleValue[82]~ AngleValue[86]: 5 DOFs of upper torso.

7) AngleValue[87]~ AngleValue[89]: 3 DOFs of lower torso.

8) AngleValue[90]~ AngleValue[95]: 6 DOFs of humanoid(Global position and orientation).

Given a vector of 96 elements, a pose can be determined, and a VRML virtual human under this pose can be created and displayed based on forward kinematics. Based on this technology, we developed a class GestureView, which can load a different VRML virtual human model and display it under any pose as shown in figure 5. This class has a convenient interface to modify the pose of the virtual human to display a human motion.

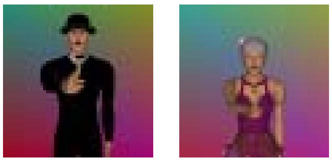


Figure 5 Chinese sign word “GREAT” demonstrated by different virtual human

4. Data mining

4.1 Hand gesture clustering

We record the motion of 5500 sign word of CLS, in other words, 10008 vectors with 21 elements that contain almost all data of hand motion.

We apply a clustering algorithm to locate the base hand states. Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ design algorithm based on a training sequence [9][10].

Clustering is a common and important method in Data Mining, but LBG clustering still have a lot of shortage: it relies on initial value, and often clusters some point that is not good. We gave an algorithm which have improved on it, added the number of initial value, this method reduce the chance of leaking clustering point, and have got a lot of good result.

As a result, 238 hand gesture classes are the centroid from the data of hand motion, each gesture is similar to one of the 238 hand gesture classes. Some hand gesture classes are shown in figure 6:

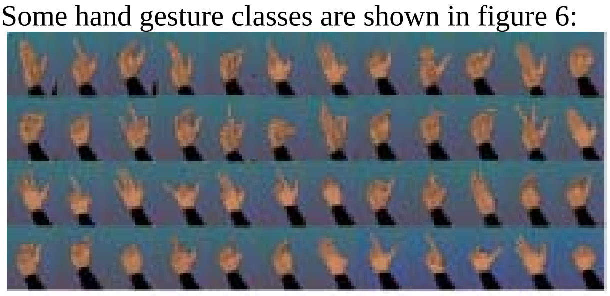


Figure 6. Some hand gesture classes

In the same hand gesture class, the hand gesture is similar. For example, the first hand gesture in figure 4 is the centroid from 1079 data of hand motion. Some gestures in the same hand gesture class are shown in figure 7:

Figure 7 Some gesture in the same hand gesture class

4.2 Rough Set and reduction

Rough set theory, introduced by Zdzislaw Pawlak in 1980s[7], is a new mathematical tool to deal with vagueness and uncertainty. One of the main advantages of rough set theory is that it does not need any preliminary or additional information about data.

The main concept of rough set theory is an indiscernibility relation. According to the concept of indiscernibility relation, it is very simple to define redundant attributes. If a set of attributes and its superset define the same indiscernibility relation, then any attribute that belongs to the superset and not to the set is redundant. Such a set of attributes, with no redundant, is called minimal. The set P of attributes is the reduction of another set Q of attributes if P is minimal and the indiscernibility relations, defined by P and Q, are the same.

We have put forward an effective reduction algorithm based on an viewpoint which rough sets regard that knowledge is a disceniblity ability, quantum the knowledge and prove the quantum is reasonable, then put the ability of classing thing as heuristic information, direction the reduction.

Human hand motion is highly articulate, but also highly constrained, which makes it difficult to model [1], and it means that some joints are redundant.

Deal the 238 hand gesture classes data with reduction algorithm, we obtain the result that 6 joints are redundant, in another word, the hand has 15 DOF independently. Each independent joint corresponds with a base gesture ξ_i as shown in figure 8. For each gesture ϕ in the hand gesture space, we will represent its parameters using a polynomial: $\phi = \sum_{i=1}^{15} \alpha_i \xi_i$

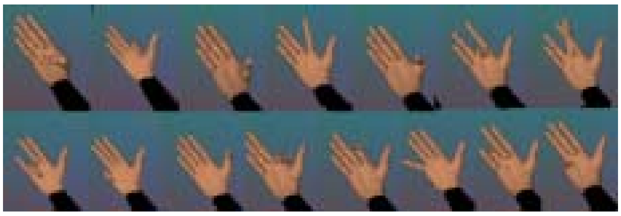


Figure 8. base hand states

5. *Conclusion*

Human hand motion is highly articulate, because the hand consists of many connected parts leading to complex kinematics. At the same time, hand motion is also highly constrained, which makes it difficult to model. Many researches use the result of hand anatomy and biomechanics, these constrains were intentionally made invalid, it is an approximation of nature finger motion.

In this paper, we presented a data mining approach to model the hand constrains. Our approach has three characteristics. First, the model is compact and believable by utilizing cluster and rough set method. Second, the hand motion data is collected by CyberGlove, which makes the result more natural. Third, the constrains that cannot be explicitly represented can be represented by reduction.

References

1 Ying Wu and Thomas S.Huang," Hand Modeling, Analysis, and Recognition". IEEE SIGNAL PROCESSING MAGAZINE, May 2001, pp.51-60

2 J.J Kuch and T.S.Huang, "Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration," in proc. IEEE Int. Conf. Computer Vision, Cambridge, MA,June 1995, pp. 666-671

3 J.Lee and T.Kunii, "Model-based analysis of hand posture," IEEE Compute. Graph. Appl.,vol.,15,no.5,pp.77-86,sept.1995.

4 J.Lin, Y.Wu, and T.S.Huang, "Modeling human hard constraint," in Proc. Workshop Motion,Dec. 2000, pp.121-126

5 T.Stamer et al., "A wearable computer based American sigh language recognizer," in Proc. IEEE Int. symp.Wearable Computing,

Oct. 1997, pp.130-137.
6 C. Vogler and D.Metaxas, “ASL recognition based on coupling between HMMs and 3D motion analysis,” in Proc. IEEE Int. Conf. Computer Vision, Mumbai, India, Jan. 1998, pp.363-369.
7 Pawlak Z. Rough Sets. International Journal of Computer and Information Science, 1982, 11(5): 341-356
8 Ken Showmake, Animating Rotation with Quaternion Curves, SIGGRAPH, 1985, 19(3):245~254.
9 S.M. Perlmutter, P.C. Cosman,C. Tseng, R.A. Olshen, R.M. Gray, K.C.P. Li, and C.J. Bergin. Image Quality in Lossy Compressed Digital Mammograms. Signal Processing. 1997,Vol. 59, No. 2, June pp. 189-210,.
10 R.M. Gray and D.L. Neuhoff. Quantization. IEEE Transactions on Information Theory. 1998, Vol. 44, pp. 2325-2384
11 VRML97 Specification, ISO/IEC DIS 14772 .
12 Humanoid Animation Specification 1.1, http://ece.uwaterloo.ca/~h-anim/ .
13