SEMANTIC FEATURE DETECTION FOR REAL-TIME IMAGE TRANSMISSION OF SIGN LANGUAGE AND FINGER SPELLING

Jin Hou and Yoshinao Aoki

Media Lab, Department of Electronics and Information, Graduate School of Engineering, Hokkaido University, N13 W8,
Sapporo 060-8628, Japan
{jinhou, aoki}@media.eng.hokudai.ac.jp

Abstract

This paper proposes a novel semantic feature detection (SFD) method for real-time image transmission of sign language and finger spelling. We extract semantic information as an interlingua from input text by natural language processing, and then transmit the semantic feature detection, which actually is a parameterized action representation, to the 3-D articulated humanoid models prepared in each client in remote locations. Once the SFD is received, the virtual human will be animated by the synthesized SFD. The experimental results based on Japanese sign langauge and Chinese sign langauge demonstrate that this algrithom is effective in real-time image delivery of sign language and finger spelling.

1. Introduction

There has been a growing interest in designing natural interactive Human-Computer Interface (HCI) that incorporates vision-based technologies, such as tracking and interpretation of body location and posture, facial expressions, and manual gesturing, which is referred to as "perceptual user interface" recently. Nonverbal language plays an important role in natural human communication. In most situations, nonverbal language is not used alone but jointly with verbal communication, strengthening or weakening sense, and it is the best and only efficient way to communicate feelings. In particular, when communication is implemented between intercultural people, nonverbal language is able to express something that would be very difficult to understand only using the linguistic system.

First attempt to natural human communication resulted in tape markers used for tracking facial expressions, and magnetic gloves used for measuring the positions of fingers and hands in real-time, allowing both face and body to be synthesized [1][2]. However, the associated measuring equipments are cumbersome that inhibit the free movement of users, and they are expensive as well. Therefore, a non-contact, non-wear method is preferable. That has spawned the more natural vision-based techniques, which are classified into appearance-based approach and 3-D model-based approach. Appearance-based approach is simpler to implement for real-time gesture recognition, but is far away from natural interface for its inherent limitations [3]. The promising 3-D model-based approach could offer more details for natural HCI, but the current systems[4][5] result in either huge computation that hinders real-time communication, or inadequate information for generating 3-D postures of hands and fingers. This failure leads to our study, which aims at an efficient delivery method for a 3-D model-based telepresence system involving detailed two-handed gestures.

In order for efficient delivery, we use an intelligent communication method: extract semantic information from the input text as an interlingua by natural language processing, and then transmit only the semantic feature detection (SFD), which actually is a set of parameterized action representation, to the 3-D articulated humanoid models prepared in each client in remote locations. Once the SFD is received, the virtual human will be animated by the synthesized SFD. Since the data volume of transferred SFD is very limited, the communication could be implemented in near real-time.

Although there are many different types of gestures, the most structured sets belong to the sign languages, so we choose Japanese Sign Language (JSL) and Chinese Sign Language (CSL) as the test bed in our pilot study.

2. Intelligent Nonverbal Communication System Overview

2.1 General Architecture of Intelligent Nonverbal Communication System

Fig.1 shows a general architecture of this proposed non-verbal communication system. Firstly, a participant inputs a natural language (Japanese or Chinese) from a keyboard (we just input the keywords for our pilot study), each keyword corresponds with a set of parameters representing actions, and we call this parameterized action representation a semantic feature detection (SFD) which is an interlingua bridging the gap between natural language input and the sign language output. Then the SFD is transmitted via the Internet. When the SFD reaches the remote participant, the intelligent decoding is done and animations of the avatar will be synthesized with the corresponding parameters. For example, when a Japanese sentence "Watasi Syuwa Benkyo Sitai (I want to learn a sign language)" is input from a keyboard, each Japanese word will be translated into a set of corresponding action parameters, by which the animations of avatars will be implemented.

2.2 Client-Server Topology

Java adds complete programming capabilities to network access, making VRML fully functional and portable.

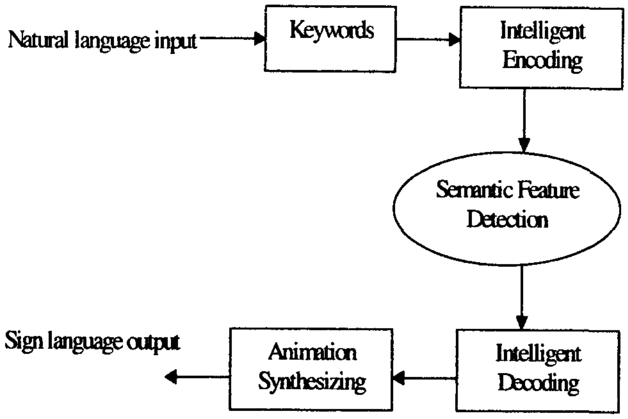

Fig.1 General architecture of intelligent non-verbal communication system

VRML/Java based shared space research started about 4 years ago [7], which supplies a new Human-Computer Interaction (HCI), where the avatars that represent the human being can communicate with each other in real time, with so much sense of presence that the user may have the illusion of being in the real world. Each person can see the animations of the avatars representing not only himself but also the others in the shared space. Each client that receives the same VRML file is connected to the same server so that the shared cyberspace is generated.

In our system, the major role of the server is to extract semantic information from the natural language input from a client and transmit it to all clients connected to the same sever. The duty of the clients is to synthesize animations of the avatar by the received parameters. The server consists of a communication object and a transformation object. The communication object arranges the list of clients. If a client sends a request such as connecting-a-link, making-a-group, or sending-a-message, the server will respond this request and distribute the message to all clients connected to each port. While the transformation object is responsible for transforming the natural language into a set of corresponding parameters or SFD according to the online digital natural language-to-parameter dictionary. While the client is composed of a user-interface object, a communication object, and an image synthesis object. The user-interface object waits for a user input such as connection command or sentence, and it displays synthesized sign language animation through the image synthesis object. The communication object sends a request or a sentence, where the corresponding response is received if that is a request or the SFD is available if that is a sentence. SFD is transferred to the image synthesis object and used to generate CG animations with the 3-D avatar.

The scenario of the communication between the server and the clients could be described by the following: (1) The server is ready for responding a request from a client. (2) A client is connected to the server. (3) The server adds the client into the existing chatting group and sends a list of online group to all clients. (4) A natural language is input from a client interface. (5) The words are translated into SFD according to the online digital dictionary. (6) SFD is transmitted to each online client via the Internet. (7) SFD is decoded and animations are implemented in the clients.

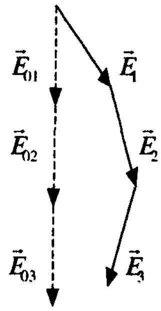

Fig.2 Movement expression of a finger based on vectors

3. Semantic Feature Detection Based On 3-D Model

3.1 Static Parameterized Representation of the Shape of Hands and Arms

The human hand skeleton consists of 27 bones based on the anatomical structure. Lee et.al developed a 27 degree of freedom (DOF) hand skeleton model. They classified the hand joints into three types: flexion or twist, directive and spherical joints which have 1 DOF (extension/flexion), 2 DOFs (one for extension/flexion and one for adduction/abduction) and 3 DOFs (rotation) respectively. Based on this theory, each finger (from index to pinky) has 4 DOFs (two at the metacarpophalangeal or MCP, one at the proximal interphalangeal or PIP, and one at the distal interphalangeal or DIP), while the thumb has five DOFs (two at the trapeziometacarpal or TM, two at the metacarpophalangeal or MCP, and one at the interphalangeal or IP). We describe these DOFs by the joint angles ϕ_{ij} (i = 0~4, j = 0~4, $\phi_{i0}=0$ when i≠0). In addition to the 6 DOFs for the position and rotation of the wrist, the hand model has 27 DOFs in total. Because the shape of hand could be decided by the joint angles of fingers, we describe the hand shape (HS) by the following formulas.

$$HS = F(\Phi) = F(\Phi_0, \Phi_1, \Phi_2, \Phi_3, \Phi_4) \quad (1)$$

$$\Phi_i = \begin{cases} (\phi_{i0}, \phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4})^T & if \ i=0 \\ (0, \phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4})^T & if \ i=1\sim4 \end{cases} \quad (2)$$

Hence, we get the parameterized hand shape representation by the joint angles of fingers. While the arm shape representation could be described by the same method. The arm shape is decided by the joint angles of arm. Since the shoulder joint has 3 DOFs, the elbow joint has 1 DOF, and the wrist joint has 6 DOFs, we suppose their joint angles are (θ_0, θ_1, θ_2), (θ_3) and (θ_4, θ_5, θ_6. θ_7, θ_8, θ_9) respectively. Therefore, the arm shape (AS) could be described with the joint angles of arm by the following formula.

$$AS = F(\Theta) = F(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9) \quad (3)$$

Therefore, we get the static parameterized representation of the shape of hands and arms.

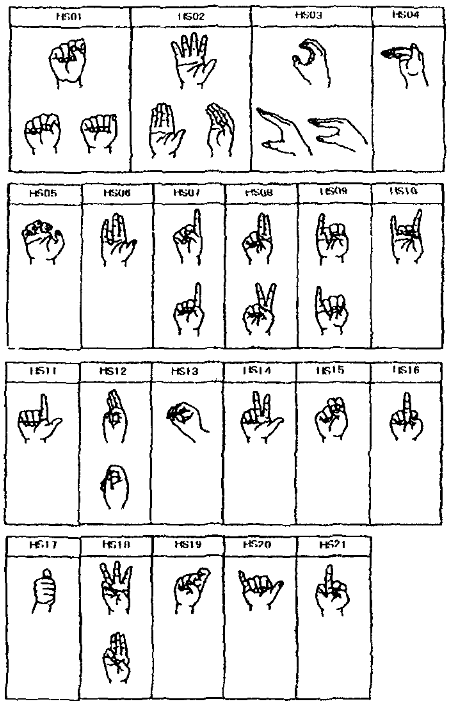

Fig.3 Hand shapes

3.2 Dynamic Parameterized Representation of the Movement of Hands and Arms

The hand and arm consist of segments and joints, and the movement of each segment is produced by rotations of its proximal joint so it can be specified by the joint rotation angles. Most calculation methods of movements are based on the dynamics analysis and neural networks, but it is not easy to express the motion of multiple joint objects such as arms and fingers precisely. In this paper, we propose a complex algorithm of vector rotation and addition. We explain this method by giving an example of a finger.

Firstly, we specify a vector for each segment of a finger as shown in Fig.2. We define local coordinate systems of the finger on every joint position with conventional right-handed coordinate systems. Each joint's rotation R is expressed by a sequence of rotations occurring around the x, y ,z axes of the local coordinate system R=Rx(α)Ry(β)Rz(γ),where α , β , γ are joint rotation angles around x,y,z respectively. Secondly, we calculate the joint motion by rotating the segment vector at the proximal joint and then adding it to the parent segment vector to get the expression in the global coordinate. This calculation is shown in Eq.(4).

$$\begin{cases} \vec{E}_1 = R_1\vec{E}_{01} \\ \vec{E}_2 = R_1(\vec{E}_{01}+R_2\vec{E}_{02}) = \vec{E}_1 + R_1R_2\vec{E}_{02} \\ \vec{E}_3 = R_1\left[\vec{E}_{01}+R_2\left(\vec{E}_{02}+R_3\vec{E}_{03}\right)\right] = \vec{E}_2 + R_1R_2R_3\vec{E}_{03} \\ R_i = R_{ix}(\alpha)R_{iy}(\beta)R_{iz}(\gamma) \qquad i=1\sim3 \end{cases} \tag{4}$$

This complex algorithm is suitable for the movement of arms as well, in that casc, thc upper arm vcctor and forcarm vector are used. Hence, we can describe the movements of hands and arms by the coordinate values of a sequence of vectors.

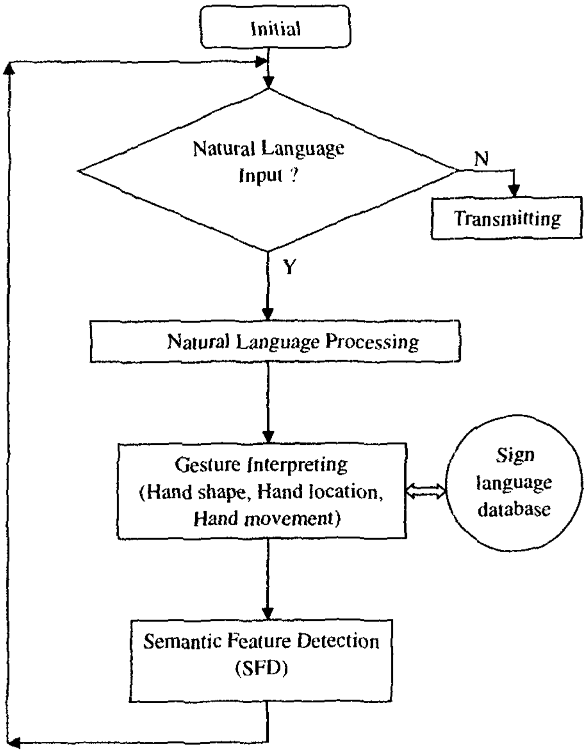
Initial
Natural Language Input ?
N
Transmitting
Y
Natural Language Processing
Gesture Interpreting (Hand shape, Hand location, Hand movement)
Sign language database
Semantic Feature Detection (SFD)

Fig.4 The flow chart of the semantic feature detection

We describe the movements of two arms and hands using their joint angle parameters together with segment lengths. Being constants, the segment lengths are sent to each client in advance, therefore, no concern is needed when synthesizing animations.

3.3 Sign Language Image Synthesis

Generally, based on the parameters of the start point and the end point of a hand, the animation is synthesized by interpolating intermediary parameters in the moving path. Supposing the parameter set of the start point is P_1 and the parameter set of the end point is P_2, and it takes (T_2-T_1) seconds from P_1 to P_2, the parameter set of any intermediary point P_t in the moving path can be calculated by the Eq. (5).

$$P_t = P_1 + \frac{P_2 - P_1}{T_2 - T_1}(t - T_1) \tag{5}$$

where t ($T_1 \le t \le T_2$) is the time when the hand reaches the intermediary point.

According to this method, more keyframes are interpolated, more precisely the animation can be described. Considering the real time animations, we use a heuristic function to estimate the in-between keyframes as shown in Eq. (6).

$$F_t = \max_{1\le j\le 62}\left\{\frac{11}{180}\left|P_{T_1}^{j} - P_t^{j}\right|\right\} + 2 \tag{6}$$

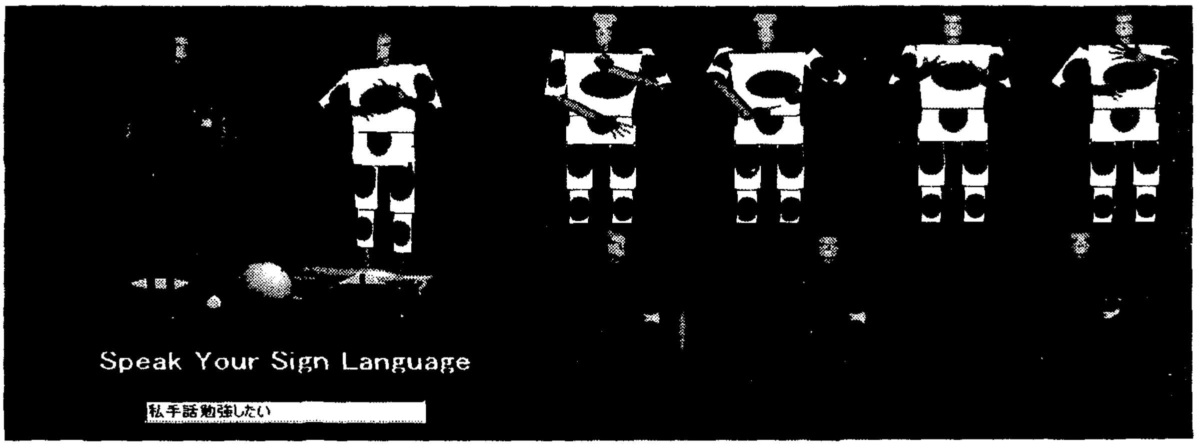

Fig.5 Sign language animations from natural language input

where Ft is the in-between keyframe number at t after T_1, j representing the parameter index, totally 62 parameters used.

3.4 Semantic Feature Detection From the Hand Shape, Hand Location and Hand Movement

We classify 21 kinds of hand shapes (coded as HS01~HS21) as shown in Fig.3, and define 30 locations (coded as HL01~HL30). A visual sign language is expressed by a sequence of actions, i.e., animations of avatars. Animations are described by the start location, the end location and the moving path. The moving path can be worked out by some interpolators between the start point and the end point. In other words, if we know the start and the end locations of a hand, we can interpolate some points to the route according to the movement of a hand, so that animations can be described very precisely. We define 18 kinds of movement (coded as HM01~HM18) for a hand such as rotation, shaking, bending, and so on.

Therefore, while we build an sign language database where some parameters could be calculated offline in advance, the parameters (SFD) for sign language real-time description could be indexed rapidly online. Fig.4 shows the flow chart of the semantic feature detection.

4. Experiment

We implement experiments of JSL and CSL with the proposed method. Fig.5 shows the result of an experiment, where the Japanese sign language animation means " I want to learn a sign language" according to the input Japanese "Watasi (Wa) Syuwa (Wo) Benkyo Sitai" (the auxiliary wordsWa, Wo are omitted), while the Chinese sign language animation means " I teach you" according to the input Chinese " Wo Jiao Ni ".

We have built an online digital sign language dictionary with about 80 corresponding natural language words registered. Currently, we have been carrying out experiments among JSL,CSL and KSL, and the recognition rate of sign languages reaches up to 84% with nearly real-time synthesizing, which is considered acceptable as an initial study. This proposed system is a promising HCI as our digital natural language-to-parameter dictionary grows.

5. Conclusion

This paper presents an efficient delivery method, semantic feature detection (SFD), for real-time image transmission of sign language and finger spelling. This method incorporates natural language processing with intelligent communication: analyse the input natural language, and then extract the semantic information data that is a sequence of parameters from the input text, and finally transmit the limited data only for animating the 3-D hierarchical avatars. This method could be extended to other sign languages than JSL and CSL.

6. Acknowledgement

This work is partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant 13305026.

7. References

[1] S.S.Fels and G.E.Hinton, "Glove-talk: a neural network interface between a data-glove and a speech synthesizer," *IEEE Trans. Neural Networks*, vol.4, pp.2-8, Jan.1993.

[2] D.J.Sturman and D.Zeltzer, " A survey of glove-based input," *IEEE Comput. Graph. Appl.*, vol.14, pp.30-39, Jan.1994.

[3] A.Lanitis, C.J.Taylor, T.F.Cootes, and T.Ahmed, "Automatic interpretation of human faces and hand gestures using flexible models," in *Proc. Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp.98-103, June 1995.

[4] K.Singh, J.Ohya, and R.Parent, " Human figure synthesis and animation for virtual space telteconferencing," in *Proc.VRAIS'95*,pp.118-126, 1995

[5] C.R.Wren, A.Azarbayejani, T.Darrell, and A.P.Pentland, " Pfinder: Real-time tracking of the human body," *IEEE Trans. PAMI*, vol.19, no.7, pp.780-785, July 1997.

[6] J. Hou and Y. Aoki, "Parameterized action representation from natural language for autonomous agents in virtual cyberspace," *Proceedings of the 2002 IEICE general conference (with CD-ROM)*, pp.284, Waseda University, Tokyo, Japan, Mar. 2002.

[7] R. Lea, K. Matsuda and K. Miyashita, *Java for 3D and VRML Worlds*. Prentice Hall, Japan, 1997.