3D Anthropomorphic Avatar Model Based on Multi-layered Representation

Zheng Wang and Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China

***Abstract*—Avatars are virtual characters making the communication between user and machine more natural and interactive. The main advantage of using avatars within interfaces is to allow user to interact intuitively with the system by giving him the illusion of communicating with a real human. This illusion is obtained by mimicking human communication, i.e. giving avatar the ability of expressing emotions through facial and body language. We proposed the method based on multi-layered representation for a kind of 3D anthropomorphic avatar, and validated it as an information presenter. The way of controlling the avatar's emotion is to use parameterized facial muscle model and skeleton skinning, scheduled with multi-thread. Then a series of naturalistic avatar animation produced.**

***Index Terms*—Avatar, Body language, expression, facial animation**

I. INTRODUCTION

From Sanskrit avatar denotes anthropomorphous Vishnu in Hinduism, cruising in human world. And now, under the influence of Snow Crash, a popular science fiction in 1990's, avatar has been authorized by customs to refer to virtual human in a virtual cyberspace or game world, especially in research field.

When all is said practical activities can't do without human, so it is very important to introduce avatar into some research topics such as virtual environment, simulation and digital entertainment. Firstly, in a virtual environment avatar is improving interactivity obviously between human and machine by inpouring vitality. Secondly, avatar can supplant human to fulfill dangerous and repetitive tasks so as to ameliorate human's life. Finally, we can acquire a deep insight into cognition of our surroundings by making full use of avatar evaluation

Recently, research topics about avatar mainly focus on the following aspects:

(1) Geometry modeling. Though there exist so many methods to construct geometry model of human body, we have to take the limited computation resource into account when a complex, realistic and accurate surface will be rendered in conditions of real time.

(2) Motion controlling. In an anatomical perspective, as a complicated being avatar can be disassembled into skeleton, muscle and skin. Thereinto, the skeleton will determine avatar's orientation, while the muscle and skin will morph avatar's surface. We can generate motion of the skeleton by key frame, motion capture, forward and inverse kinematics, and dynamics method. Generally, the above methods should be resorted to in a form of combination to achieve a maximum effect.

(3) Intelligent behaving. It is increasingly becoming a hot spot to model avatar's intelligence to simulate autonomous behavior of human being in a virtual environment based on geometry and motion model these days.

It is well known that as far as speech is concerned, human's behavior is a multi-modal form of communication, seeing the face and body language (BL) of a talker provide information that can significantly influence the perception and understanding of his intention[1]. Therefore, it is reasonable to focus attention on the visual modality in addition to the auditory in the synthesis of speech based on traditional agent theory framework and the above mentioned geometry and motion model. The system we proposed will make an emphasis on the visual modality of intelligent interaction between avatar and persons.

The paper is organized as follows. An overview of previous work similar with ours is described in Section II, and some differences or feature points will be highlighted too. Section III presents the system's architecture, and relative key technologies will be expounded in Section IV. Section V presents a typical scenario as an application and experimental evaluation. Finally, in Section VI, we make some conclusions and comments.

II. OVERVIEW

2.1 Previous work

The most common implementations of this kind of avatar consist mainly of talking heads which use some variations of TTS (text-to-speech) and dialog management software to drive an intelligent conversation with user or to present some information in an oral way [2][3]. Examples of talking heads which help in delivering their web-site's information include the Ananova web site, http://www.ananova.com, which

Manuscript Received on 9 July, 2010
E-Mail:flair@126.com

features a video news report with a computer generated face animation using TTS synthesis. Talking heads also have been developed to be counselors for helping to eliminate smoking habit[4]. Many cell phone applications are available: concierge services like news, horoscopes, weather, the nearest restaurant, and sports, and virtual teacher[5]. Moreover, there have been several educational applications: helping develop competencies in inquiry, analysis and synthesis[6]; language tutoring for children with hearing loss[7]; and teaching to write and read[8]. Another similar application used to present and search contents can be found on the KurzweilAI web site, http://www.kurzweilai.net. This site features a more dialog-oriented interactive talking head. Even though the interactive level is higher, the avatar expressiveness is limited to facial gestures. We can also verify many limitations in the areas of speed and visual appearance, intelligent dialog management and interactivity from other examples of the talking head approach: http://www.extempo.com, http://www.sensoryinc.com, http://xface.itc.it/.

Among the areas to improve, a very important missing detail is the visual modality to show full body expressions and gestures. We propose to extend the talking-head model to a full-body virtual human which will be able to complement TTS voice synthesis, dialog management and facial animation with body gestures to enhance the expression and give a more human touch to avatar through use of non-verbal communication [9] [10] [11].

A typical example, named Nadia, was set up. Nadia is demonstrated at http://clone3d.net, and it is a cartoon-styled avatar able to perform dialogues with users by ALICE chatbot, generating English phonemes with automatic lip-sync, and expressing simple emotion, including body movements, hand actions, and facial gestures. The lighting of Nadia is practically naturalistic and uses conventional illumination techniques. To a great extent, Nadia's function and performance is similar with Maxine developed by Sandra Baldassarri[12], and Amalia Ortiz's work based on VHML[13].

2.2 Our work

Interpersonal communication is a complex phenomenon that occurs in a specific context. Littlejohn[14] suggested five conditions must exist. First, there must be two or more people in close immediacy who can sense the presence of each other. Second, there is interdependence in which the interaction is affected by one another. Third, there should be exchange of messages. The forth one is concern with the coding of messages in two different channels: verbal and non-verbal, and the last one is establishing the exchange with two features: flexibility and informality [15].

The conversation with an avatar is an emulation of interpersonal communication. Researchers have been addressing some of the conditions needed in the interaction between the avatars and people. In a field study, presence was investigated as the first condition. The results of the study indicate even limited copresence, the degree to which a user judges she is not alone and isolated, supplied by a prototype avatar is satisfactory to facilitate users to experience presence. Perception of sensory stimuli and the understanding of symbols In this paper we presented a new set of tools designed to are ways to sense the others' presence. In this case, presence is closely related to immersion [16].

improve the visual modal HCI (human-computer interaction) with immersion by means of multi-layered representation methodology. The developed software is a real-time multiple platform, aimed to produce realistic full-body and facial animation of 3D avatar in a human-like style on diversified applications.

One of the most important features we tried to include in the developed platform was the capability to synthesize full-body animation instead of limiting it to face animation. Using body gestures, the avatar will be able to express natural human reactions, such as emotions, giving the illusion that the user is interacting with a real person inside the computer. While using facial gestures, the avatar can produce about 60 expressions such as pleasure, anger, sorrow, joy and others, so the user will feel much more immersive.

In the next section we will justify the innovations of our platform and relative key technologies.

III. ARCHITECTURE

Our system is built to be modular and user-extensible, and to operate in real-time. To this end, it is written in C++, based on an input-to-output POSIX multi-thread approach with support of user defined filters and knowledge bases, and uses XML as its primary data structure. Processes are decomposed into threads which operate as XML transducers; each taking an XML object tree as input and producing a modified XML tree as output. The first module in the process operates by reading in XML-tagged text representing the text of the avatar's script and converting it into a parse tree. The various knowledge bases used in the system are also encoded in XML so that they can be easily extended for new applications.

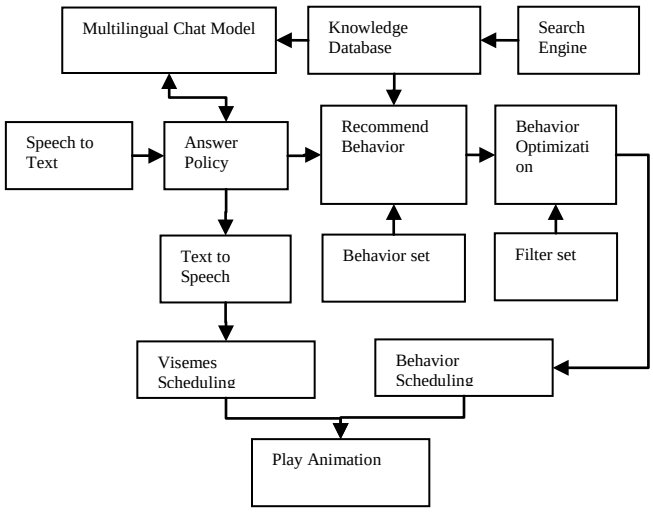

Fig.1. System Architecture

From the perspective of running procedure, the overall architecture of our system is shown in Fig. 1. The user formulates any order, question or sentence that might be used in conversation in natural language. The sound or audio generated

by the user is picked up by microphone and sound card. One of the main requisites of our system is that it must be able to understand and speak Chinese. This constraint prevented us from using the existing open source libraries, all of which are in English. Therefore, in order to obtain a text chain from the words delivered in Chinese by the user, a voice recognition engine and TTS engine were built by another group in our institute in the past 2 years. The Behavior Recommendation and Optimization module mainly derived from annotation schemes for conversational gestures.

In respect to visual modality of the avatar, the system has two main input files as well as one output file, the whole configuration file defining all of the interfaces among modules, the behavior file that define the geometry of avatar, which includes the information for anatomic deformations, and body animation parameters which contain information to alter the avatar links and display animation, and an output file from Phoneme and Behavior Scheduling module defining external control protocol by a simple XML description.

IV. KEY TECHNOLOGIES

During the development of the system, special attention was paid to creating immersive visual modal interaction via multilingual text and speech. This broadens the spectrum of potential users of the system, for example, English and Chinese speakers at different ages with different levels of education, the hearing-impaired or paraplegics, and people with or without computer knowledge. With the ultimate aim of enhancing interaction and establishing emotional communication between user and avatar, it is essential to describe the involved key technologies.

4.1 Multilingual input/output

TABLE 1: D EFINITION OF BASIC VISEMES USED BY AVATAR

Initial		Final		
b/p/m (b/p/m)	g/k/h (h/k/g/ng)	a/ang (aa)	ou (ow)	i (y/iy/ih/ix)
f (f/v)	j/q/x (th/dh)	ai/an (ey)	e/eng (eh)	u (w/uw)
d/t/n (d/t/n)	zh/ch/sh/r (zh/ch/sh/r)	ao (ao/aw/oy/ay)	ei/en (ae/ax/ah)	v/ü (jh)
l (l)	z/c/s (z/s)	o (uh)	er (er)	neutral

ao er

Fig. 2 Visualized visemes

As mentioned above, we implemented a multilingual TTS based on similarity between English and Chinese Mandarin phoneme, a monosyllabic language/dialect where each syllable can be divided into an optional initial (i.e. a consonant) and final (i.e. a core vowel followed by an optional consonant). There are 21 initials and 38 finals in Mandarin, some of which share the same viseme with 21 English visemes. Hence, this work defines 20 basic visemes in all. In Table 1, initials/finals sharing the same viseme are grouped together. There is a viseme for "neutral", i.e. the natural state of the face without speaking and expression. Corresponding to these visemes, avatar's performances are shown in Fig. 2.

4.2 Multi-layered representation

In terms of different function of avatar's representation during implementing, we have 3 functional parts including head, body and accessories. The head will transfer expression and emotion, dialog with user, while the body will perform behavior language and accessories will decorate sense of sight.

In terms of different domain model, we have geometry, motion and render models. Each model consists of some sub-components. For example, a head geometry is assembled into a unit with hair, tooth, mouthpiece, face, eyelids and eyeballs etc, and a motion model is composed of body behavior based on skeleton skinning, expression and lips behavior based on Keith Waters' muscle theory.

There is a special reason in processing hair and cloth accessories, because hair can be modeled with polygons or splines, at the same time mass-spring physics model can be partitioned into static and dynamic sections for hair and cloth to accelerate FPS.

4.3 Expression modeling

Because morphing targets need to be redone if you change the number of vertices in a mesh, but muscles have no this kind of limits. Referring to Keith Waters' muscle model, we have defined 26 embedded muscle vectors on avatar's face in symmetry to generate deformation units in FACS.

According to anatomy, for most of faces linear muscle vector starts from point V_1 attached on bones to V_2 embedded into soft tissue. Given muscles deform with equal stretch or contraction force, the vector can be represented with direction and length. Fig.3 describes an impact on point P made by neighbor vectors, as can be extended to the case of 3D space. If V_2 has a maximum displacement while V_1 has a minimum displacement, we can calculate the displacement $P'(x',y',z')$ of one point $P(x,y,z)$ to simulate muscular motion with nonlinear interpolation method, and R_s, R_f for start point and end point of attenuation radius, respectively. In the domain of $V_1P_rP_s$, P displaces in direction of PV_1, so we have

$$x' \propto f(K \cdot A \cdot R \cdot x) \text{ and } y' \propto f(K \cdot A \cdot R \cdot y) \quad (1)$$

where K is muscle spring coefficient, Ω is an angle from the max domain of influence, and $D = ||V_1P||$.

The angular displacement factor is defined as follow,

$$A = \cos(\mu/\pi \cdot \pi/2) \quad (2)$$

where μ is the angle between V_1V_2 and V_1P. The radial displacement factor is defined as follow,

$R = \cos((1-D/Rs)\pi/2)$, if P is in the domain $V_1P_nP_m$,

$R = \cos((D-R_s)/(R_s-R_f)\pi/2)$, if P is in the domain $P_nP_rP_sP_m$.

The jaw and eyelids rotating angle θ refers to axis x, obtained as follow,

$$x' = x \quad (3)$$

$$y' = y\cos\theta - z\sin\theta \quad (4)$$

$$z' = y\sin\theta + z\cos\theta. \quad (5)$$

The effect of the above deformation algorithm can be seen in Fig. 4.

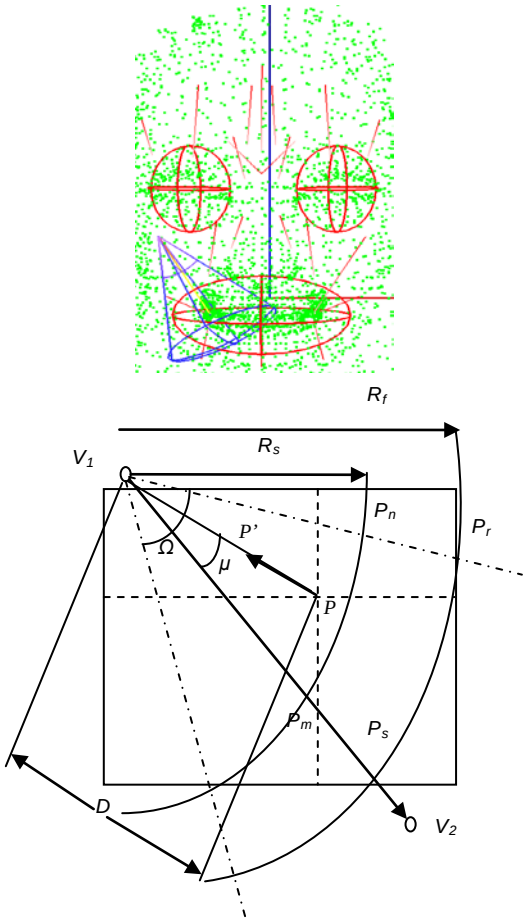

Fig. 3 Muscle vector in the domain $V_1P_rP_s$

surprise laugh

Fig. 4 Visualized expressions

4.4 BL modeling

BL highlights the attitude of avatar towards what she is saying and comments on the relationship of avatar and addressee, which is typical of metaphorical and beat gestures. For example, a gesture in which the avatar brings her hands toward her head and shakes them to pantomime a frustrated reaction should count as BL. So should a gesture in which the avatar points at the addressee and shakes her hand side-to-side in reprimand. BL should be indexed under the function associated with the performance. Pantomimed frustration will fit any context that reports a frustrating outcome. An embodied reprimand fits any context where the avatar describes something the user has done wrong.

The interdependence between function and content of BL is very close as part of annotation, because of the complexity of BL semantics. It is especially hard to label BL with a specific communicative function and then reason correctly about it. We assume that the content and function that the template generator associates with each performance unit already characterizes BL precisely enough for the application.

We use performance-driven animation technology by recording human motion and rendering it back just as performed. Performance data can also serve as the basis for synthesizing new motion by warping captured motion in time and space, interpolating captured motion to vary its emotional avatar, retargeting it to new avatars, and preserving its dynamics in the process. This kind of manipulation remains limited in its ability to adapt the gross structure of captured motion. So we use a more extensive database of captured motion to achieve more flexible synthesis. In the database, captured motion data will be segmented into short stored units as behavior set that can be blended together for behavior recommendation. The approach constructs new motions by selecting sequences of units to splice together so as to optimize the transitions between them and to satisfy global constraints on motion.

Taking generic property into account, we adopt Physique Envelop of Biped as a skeletal template and all animation sets will be processed link by link, as in Fig.5. We use flexible skinning algorithm for vertices in mesh, where a vertex position will be determined by all of its neighbor links, and in an updated keyframe the vertex position V' in *WCS* (world coordinate system) can be calculated from previous V as follow,

$$V' = w_1M_1L_{1\text{-}1}V + w_2M_2L_{2\text{-}1}V + \ldots + w_nM_nL_{n\text{-}1}V \quad (6)$$

$$w_1 + w_2 + \ldots + w_n = 1 \quad (7)$$

where M_j is the transformation matrix of motion data, L_j is the transformation matrix of *LCS* (local coordinate system) to *WCS* for *link j*, w_j is the impact factor for vertices by *link j*.

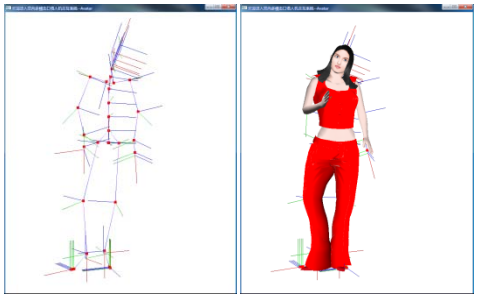

Fig. 5 Physique Envelop and BL units

4.5 Expression modeling

A key concern for scheduling is the range of behaviors that need to tightly synchronized with speech, including not only behaviors such as visemes that represent the physical manifestation of the speech but also a range of BL such as gestures, head movements and eye brow movements that can be used to augment, emphasize or replace aspects of the spoken dialog. In general, there are two ways to achieve synchronization between an avatar animation subsystem and a subsystem for producing the avatar's speech (either through a TTS or from recorded audio samples). The first is to obtain estimates of word and viseme timings and construct an animation schedule prior to execution. The second approach is to assume the availability of real-time events from a TTS-generated while the TTS is actually producing audio and

compile a set of event-triggered rules to govern BL generation. The first approach must be used for recorded audio-based animation or TTS engine, while the second must be used with TTS. We have used both approaches in our system, capable of producing both kinds of animation schedules.

The first step in time-based scheduling is to extract only the text and intonation commands, translate these into a format for TTS, and issue a request for word and viseme timings. In our implementation, the TTS runs as a separate process. Thus part of the scheduling can continue while these timings are being computed. However, If one considers the real case of human beings speaking very quickly, it is impossible to read their lips because there are visemes that last for a very short time or do not take place at all. The improvement detects visemes that do not last long enough to be reproduced because the frequency with which they are checked is insufficient. These visemes will be discarded, thus improving synchronization between the animation and the audio.

The next step in the scheduling process is to extract all of the BL optimization from database, translate them into an intermediate form of animation command, and order them by word index into a linear animation proto-schedule. Once the word and viseme timings become available, the proto-schedule can be instantiated by mapping the word indices into execution times (relative to the start of the schedule). The schedule can then also be augmented with facial expression commands to lip-sync the visemes returned from the TTS.

The final step of scheduling involves compiling the abstract animation schedule into a set of legal commands for whichever animation subsystem is being used. This final compilation step has also been modularized in the system. In addition to simply translating commands it must concern itself with issues such as enabling, initializing and disabling different animation subsystem features, gesture approach, duration and relax times (the abstract schedule specifies only the peak time at start of phrase and the end of phrase relax time), and any time offsets between the speech production and animation subsystems.

V. APPLICATION

The described constructing solution of avatar has been integrated in a demonstration application developed in the framework of a live help service project with support of multilingual TTS. The demonstration is an Internet-based application whose main objective is to synthesize human body and facial gestures corresponding to emotions. The idea of the demonstration is that anthropomorphic avatar can preserve multimodal communications and conventional conversational habits of face-to-face interaction in a live help service. Obviously, the use of nonverbal communication capabilities, facial expressions, hand gestures, and body postures can easily be perceived and understood by clients and, at the same time, enrich their interactive experiences. There is a simple framework description about this project in Fig. 6 and Fig. 7.

Fig. 6 Avatars as translator and chatbot for information presenter (Color Plate 5)

Our avatars have a degraded level of detail at the request of real-time rendering, but they can emulate natural protocols just enough to achieve recognition of familiar features, like a smile, a waving hand and a nodding head. And we can easily get about 30 FPS on a standard PC hardware platform.

In this demonstration, interdependence has been conceptualized as scheduling process, initiation, turn-taking, feedback, and breaking away. The exchange of messages among users and avatar has been addressed in different way: creating specific domain knowledge and creating models that facilitate the communicative interchange, such as discourse recipe-based model. This model uses both discourse recipes and reasoning in the context pilot communication. Discourse recipes exploit avatar's experience by encoding recurring dialogue structures.

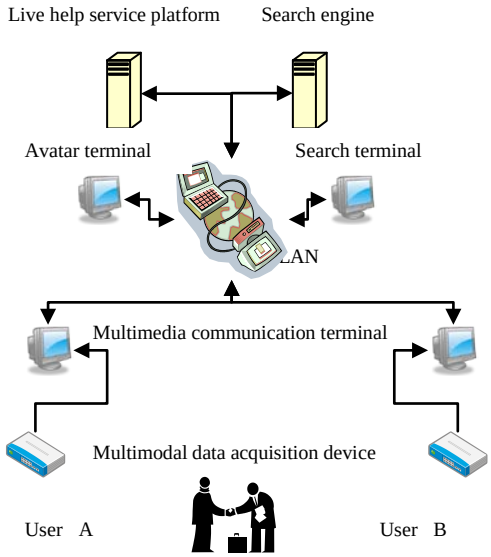

Fig. 7 Framework of the live help service project

VI. CONCLUSIONS

In this paper we have presented a real-time naturalistic avatar with multi-layered representation methodology, integrated as a user interface in an Internet-based application for a live help service with the support of a multilingual TTS engine. Our method provides a new way to implement a user interface by means of an avatar that can establish non-verbal communication using body gestures and work in coordination with TTS, dialog managers and other interaction methods as well.

The proposed system for multilingual lip synchronization is suitable for real-time and offline applications. Expression and visual representation of phonemes, visemes, defined by muscle model, are used for face synthesis. Database used for BL units is retrieved by indices from annotation and blended for behavior optimization. Speech is first segmented into frames. For each frame most probable viseme is determined. Classification of

speech into viseme classes is performed. Then facial animation is produced. Finally, by use of scheduling process, a naturalistic avatar will be rendered in a standard PC at real-time.

In application, social cues like face, BL and voice of the avatar motivate interpretation that the conversation exchanged with the avatar is similar to one with humans.

ACKNOWLEDGMENT

The authors would like to thank all persons who took part in perceptual tests and Deqiang Hu, Meng Meng for their help in mapping the output from the synthesizer to the behavior of avatar. The authors are also grateful to National Natural Science Foundation of China (No. 90820303/F030511) in carrying out this research.

REFERENCES

[1] D.G. Stork, M.E. Hennecke,. Speechreading by Humans and Machines: *Models, Systems and Applications*. 1996, Vol.150. Springer-Verlag, Berlin.

[2] C.Busso, Z. Deng, M. Grimm, U. Neumann, S. Narayan, Rigid head motion in expressive speech animation: analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 2007, 1075-1086.

[3] E. Cosatto, J. Ostermann, H.P. Graf, J. Schroeter,. Lifelike talking faces for interactive services. *Proceedings of the IEEE 91 (9)*, 2003,1406–1429.

[4] O. V.Mourik,. Can't quit? Go online. *Psychology Today*, 2006, 39(5): 27-28.

[5] E. Schwartz, Flash apps forphones. *InfoWorld*, 2005,27(20), 21.

[6] A.C. Graesser, D.S. McNamara, K. VanLehn,. Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, 2005, 40(4), 225-234.

[7] D.W. Massaro, M.M. Cohen, J. Beskow, S. Daniel, R.A. Cole,. Developing and evaluating conversational agents. *Embodied conversational agents*. 2001, 287-318.

[8] K. Ryokai, C. Vaucelle, J. Cassell, Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning*. 2003, 19, 195-208.

[9] N. Grammalidis, G. Goussis, G. Troufakos, M.G. Strintzis, Estimating body animation parameters from depth images using analysis by synthesis. *Proceedings of the Second International Workshop on Digital and Computational Video*, 2001.

[10] E.C. Prakash, A human touch for Internet training: the Virtual Professor. *TENCON '99. Proceedings of the IEEE Region 10 Conference* ,Vol(2) , 1999.

[11] G. Todesco, R.B. Araujo, MPEG-4 support to multi-user virtual environments. *Proceedings of the 20th International Conference on Distributed Computing Systems*, 2000.

[12] Sandra Baldassarri, Eva Cerezo, Francisco J. Seron. Maxine: A platform for embodied animated agents. *Computers & Graphics*, 2008(32), 430-437.

[13] Maria del PuyCarretero, David Oyarzun, Amalia Ortiz. Virtual characters facial and body animation through the edition and interpretation of mark-up languages. *Computers & Graphics*, 2005(29), 189-194.

[14] S. W. Littlejohn, Theories of human communication. Belmont, CA, *Wadsworth Pub*. Co. 1983.

[15] L. Gong, How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behaviour*, 2008,1494-1509.

[16] M. Gerhard, D. Moore, D. Hobbs. Close encounters of the Virtual kind: Agents simulating copresence. *Applied Artificial lntelligence*, 2005,19(4), 393-412.

Zheng Wang is an assistant professor at the Institute of Automation of Chinese Academy of Sciences. His area of research is centered on virtual reality, CAD and dynamics simulation.

Bo Xu is a professor at the Institute of Automation of Chinese Academy of Sciences. His area of research is centered on natural language process and digital content process, and these research have been supported by Chinese government fund including National Science Foundation and 863 High-tech plan.