Human-virtual human interaction by upper body gesture understanding

Xiao, Yang; Yuan, Junsong; Thalmann, Daniel

2013

Xiao, Y., Yuan, J., & Thalmann, D., (2013). Human-Virtual Human Interaction by Upper Body Gesture Understanding. Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology. 133-142.

<https://hdl.handle.net/10356/100286>

<https://doi.org/10.1145/2503713.2503727>

© 2013 Association for Computing Machinery This is the author created version of a work that has been peer reviewed and accepted for publication by Human-Virtual Human Interaction by Upper Body Gesture Understanding , Association for Computing Machinery. It incorporates referee' s comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: <http://dx.doi.org/10.1145/2503713.2503727>].

Downloaded on 26 Apr 2023 19:43:38 SGT

Human-Virtual Human Interaction by Upper Body Gesture Understanding

Yang Xiao*

Inst. for Media Innovation
Nanyang Tech. Univ., Singapore

Junsong Yuan†

Sch. of Electrical and Electronic Engineering
Nanyang Tech. Univ., Singapore

Daniel Thalmann‡

Inst. for Media Innovation
Nanyang Tech. Univ., Singapore

Abstract

In this paper, a novel human-virtual human interaction system is proposed. This system supports a real human to communicate with a virtual human using natural body language. Meanwhile, the virtual human is capable of understanding the meaning of human upper body gestures and reacting with its own personality by the means of body action, facial expression and verbal language simultaneously. In total, 11 human upper body gestures with and without human-object interaction are currently involved in the system. They can be characterized by human head, hand and arm posture. In our system implementation, the wearable Immersion CyberGlove II is used to capture the hand posture and the vision-based Microsoft Kinect takes charge of capturing the head and arm posture. This is a new sensor solution for human-gesture capture, and can be regarded as the most important contribution of this paper. Based on the posture data from the CyberGlove II and the Kinect, an effective and real-time human gesture recognition algorithm is also proposed. To verify the effectiveness of the gesture recognition method, we build a human gesture sample dataset. Additionally, the experiments demonstrate that our algorithm can recognize human gestures with high accuracy in real time.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality I.5.4 [Pattern Recognition]: Applications—Computer vision;

Keywords: gesture understanding, virtual human, interaction, the Immersion CyberGlove II, the Microsoft Kinect

Links: DL PDF

1 Introduction

In virtual reality, virtual human is an essential part that is widely used in game [Rautaray and Agrawal 2011], virtual training system [Wang et al. 2012a], virtual diagnosis system [Rizzo et al. 2011] and virtual user guide system [Swartout et al. 2010], etc. For these applications, how to realize the communication and interaction between human and virtual human is a crucial problem. A lot of previous studies [Gratch et al. 2007; Krämer et al. 2003; Swartout et al. 2010] have demonstrated that humans prefer to interact with the virtual human in the same way as they do with the

real humans. Thus, the human-virtual human interaction should be natural and intuitive like the interaction between real humans.

Humans communicate and interact with each other using both verbal and non-verbal languages. Verbal language is one of the most natural forms of human-human interaction. And it has been used in many virtual systems [Gratch et al. 2007; Stedmon et al. 2011; Swartout et al. 2010] for human-virtual human interaction. However, speech recognition accuracy is likely to be affected by background noise, human accents and device performance. Learning and interpreting the subtle rules of syntax and grammar in speech is also a difficult task. These negative factors limit the practical use of verbal language. On the other hand, non-verbal communication constitutes nearly two-thirds of all communication between humans [Gobron et al. 2012]. Thus, natural non-verbal language can be employed as another human-virtual human communication way to enhance the interaction performance. However, to our knowledge, few works pay attention to this topic in virtual reality community.

In this paper, we have undertaken this effort to make use of non-verbal language for human-virtual human interaction. Non-verbal language mainly consists of gaze, facial expression and body gesture [Gobron et al. 2012]. Our work focuses on understanding the meaning of human upper body gestures which could be characterized by head, hand and arm posture. The gestures being involved possess intuitive semantics for human-virtual human communication. And the gestures with *human-object* interaction are also considered. It is worth noting that the gestures with human-object interaction are much more difficult to recognize than the ones without such interaction. To bridge human and virtual human for natural interaction, a gesture understanding and virtual human interaction (GUVHI) system is implemented. So far, this system is an experimental prototype that can understand 11 human upper body gestures and leading the virtual human to react accordingly.

The major challenge to constructing the GUVHI system is how to recognize the human upper body gestures accurately and in real time. To achieve this goal, two crucial problems need to be solved:

- *First, suitable human gesture-capture devices should be chosen.* For our application, both hand posture and upper body posture are needed for gesture understanding. Because the vision-based sensors (such as the RGB camera) are difficult to capture the hand posture robustly [Lu et al. 2012; Teleb and Chang 2012], the wearable CyberGlove II [Immersion 2010] is employed. Using this device, high-accuracy hand posture data can be obtained stably. On the other hand, the Microsoft Kinect [Shotton et al. 2011] is a recently emerged low-cost depth sensor that is successfully applied to human body tracking. The body skeletons can be extracted from the Kinect depth images [Shotton et al. 2011]. In our work, Kinect is used to capture the upper body posture;
- *Secondly, effective and fast gesture recognition algorithm is required.* Based on the CyberGlove II and the Kinect posture data, descriptive upper body gesture feature is proposed by us. To enhance the performance, LMNN distance metric learning method [Weinberger and Saul 2009] is applied. Finally, an energy-based classifier is used to make the recognition.

*e-mail: xiaoyang@ntu.edu.sg

†e-mail: JSYUAN@ntu.edu.sg

‡e-mail: danielthalmann@ntu.edu.sg

In order to evaluate the proposed gesture recognition algorithm, we construct a human upper body gesture dataset for experiment. This dataset contains gesture samples from 23 people. They are of different genders, body sizes and races. Both the CyberGlove II and the Kinect data are provided. The experiment demonstrates the effectiveness of our gesture recognition method both on classification accuracy and time consumption.

In the GUVHI system, the virtual human can respond to human body language in terms of body action, facial expression and verbal language simultaneously. It targets at providing the users with an immersive experience like human-human interaction.

Overall, the contributions of this paper include:

- An original human-gesture capture sensor solution. The CyberGlove II and the Kinect are uniquely integrated to capture human head, hand and arm posture simultaneously;
- The GUVHI system: a novel human upper body gesture understanding and interaction system. Human can communicate and interact with virtual human through non-verbal body language in this system;
- An effective and real-time human upper body gesture recognition algorithm is proposed.

The remaining of this paper is organized as follows. The related work is discussed in Sec. 2. Sec. 3 gives an overview on the GUVHI system. The human upper body gesture understanding and recognition method is illustrated in Sec. 4. Sec. 5 introduces how virtual human interacts with human. Experiment and discussion are given in Sec. 6. Sec. 7 concludes the whole paper with discussion on future research.

2 Related Work

Among the human-virtual human interaction applications, most work [Gratch et al. 2007; Stedmon et al. 2011; Swartout et al. 2010] chooses verbal language as the communication tool. As discussed in Sec. 1, verbal language still suffer from some restrictions in practical use. For example, in the museum virtual guide system [Swartout et al. 2010], the visitors cannot speak to the virtual guide directly. Their questions are posed to virtual human by the museum staff in a uniform format to ensure the speech recognition accuracy. However, this interaction way largely restricts the visitors' participation and interest. Compared to the previous systems, one important contribution of our work is to lead virtual human to interact with human by understanding the meaning of human gestures.

Human gesture recognition plays an important role in human-computer interaction. According to the sensor type, the gesture recognition systems can be categorized as encumbered and unencumbered ones [Berman and Stern 2012]. Generally, the encumbered systems require the users to wear physical assistive devices such as infrared responders, hand markers or data gloves. These systems are of high precision and fast response speed, and robust to environment changes. Many encumbered systems have been proposed. For instance, two education systems [Adamo-Villani et al. 2007] were built for deaf by using data gloves and optical motion capture devices; Lu et al. [Lu et al. 2012] proposed an immersive virtual object manipulation system based on two data gloves and a hybrid ultrasonic tracking system. Although most commercialized gesture-capture devices are currently encumbered, unencumbered systems are expected as the future choice, especially the vision-based systems. With the emergence of low-cost 3D vision sensors, the application of such devices becomes a very hot topic in both research and commercial fields. One of the most famous examples is

the Microsoft Kinect, it has been successfully employed in human body tracking [Shotton et al. 2011] and activity recognition [Wang et al. 2012b]. However, accurate and robust hand posture capture is still a difficult task for the vision-based sensors.

As discussed above, both encumbered and unencumbered sensors possess their intrinsic advantages and drawbacks. For the specific applications, they can be complementary. In the GUVHI system a tradeoff between the two kinds of sensors is made, that is the fine hand posture is captured by the encumbered device (the CyberGlove II), while the rough upper body posture is handled by the unencumbered sensor (the Kinect).

Although human gesture recognition has been widely studied, very few works apply it to natural human-virtual human interaction. Piplica et al. [Piplica et al. 2012] proposed the first system which treated the human full body gesture as the input to create interactive narratives with an AI improviser. The AI improviser could interpret the human gestures and understand how the human contributed to the narratives. The human gestures are represented by the joint information extracted from the Kinect. In this system, hand gesture was ignored and some predefined gestures were not natural. Zhao et al. [Zhao et al. 2012] suggested to control virtual human both by speech and hand gesture recognition. In this work, virtual human did not equally interact with human. It just acted according to the human's commands. Rautaray et al. [Rautaray and Agrawal 2011] made vision-based hand gesture recognition to lead virtual human to act in game accordingly. Virtual human interacted with the virtual objects obeying the human hand commands. In [Wang et al. 2012a], virtual human motion was controlled by mapping human motion onto it directly. Human motion was depicted by the joint data extracted from the depth images.

Our research is different from the previous work listed above. In the GUVHI system, only natural human gestures are used. Meanwhile, virtual human reacts to human gestures with its own personality and it does not play just as an avatar.

3 System Overview

The proposed GUVHI system aims to capture and understand the meaning of human upper body gestures and trigger the virtual human's reaction in real time accordingly. As shown in Fig. 1, the GUVHI system is mainly composed of two modules. One is the *human gesture understanding module* that works as the interface between the real world and the virtual world, and the other is the *virtual human interaction module* designed to control the virtual human's behavior for interaction. The whole system is constructed based on the client-server architecture. The two functional modules are running as clients and connected by a server. The server takes charge of the message transmission between the two clients. At this stage, our system only supports the interaction between one human and one virtual human.

A right hand CyberGlove II and a Microsoft Kinect are employed as the input sensors for the GUVHI system. Now, we mainly focus on understanding the body gestures correlated with human's right hand and right arm. In the human gesture understanding module, we uniquely combine the CyberGlove and the Kinect for upper body gesture capture and understanding, which is different from all the approaches introduced in Sec. 2. The CyberGlove is used to capture the hand posture and the Kinect is applied to acquiring the 3D position information of the human skeleton joints (including head, shoulder, limb and hand). Besides the CyberGlove, the users do not need to wear any other device. Therefore, our solution will not yield a heavy load on the users to make them feel uncomfortable. And because the CyberGlove II is wireless connected to the system, the users can move their hands freely. What is more, by fusing

Figure 1: The GUVHI system architecture.

the information from the two kinds of sensors, the GUVHI system maintains the capacity of recognizing a variety of human upper body gestures, even the ones accompanied with human-object interaction, such as “call”, “drink”, “read” and “write”. Such kinds of gestures are ignored by the previous systems. However, they frequently happen in the daily communication or interaction between real humans, which may change human’s behavior abruptly. Therefore, they should be included as an essential part for the natural human-virtual human interaction. In our system, virtual human is able to recognize and give meaningful response to such kinds of gestures. Actually, this could refine the virtual human to be of more real human characteristics.

At the gesture understanding phase, the original data from the CyberGlove and the Kinect is synchronized firstly. After that, descriptive features are extracted from them respectively. The multimodal features will lastly be fused together according to certain rules to generate the uniform input for the upper body gesture classifier. The output of the classifier is regarded as the gesture understanding result and sent to the virtual human interaction module via server to trigger the virtual human’s reaction.

The virtual human interaction module enables virtual human to respond to the real human’s body gesture language. In our system, virtual human’s behavior is composed of three parts: body action, facial expression and verbal language. Combining these behavior parts concordantly can make the virtual human more lifelike, and it is also helpful to raise the users’ interest to engage in the interaction with virtual human.

As shown in Fig. 2, when the GUVHI system is running, the virtual human will be projected on a big screen and the user worn the CyberGlove is assigned to stand facing the screen in the distance of three or four feet. And the Kinect is placed besides the big screen to the capture the human body skeleton information.

4 Human Upper Body Gesture Understanding

It can be seen from Sec. 3 that human upper body gesture understanding plays a key role in the GUVHI system. Its performance will affect the interaction between human and virtual human greatly. In this section, our upper body gesture understanding method by using the CyberGlove and the Kinect will be illustrated in details. First, we introduce the upper body gestures that need to be understood in the system. The feature extraction approaches for both the CyberGlove and the Kinect are then presented. To form an integral

Figure 2: The GUVHI system deployment.

description on the upper body gestures, the features extracted from the two sensors will be fused as an uniform input for the classifier. Aiming to enhance the gesture recognition accuracy, LMNN distance metric learning method is applied to mining the best distance measures. The energy-based classifier is used to make the final decision.

4.1 Gestures in the GUVHI System

Currently, 11 static human upper body gestures can be understood by the GUVHI system. Because only one right hand CyberGlove is involved in the system, all the gestures are mainly triggered by the human’s right hand and right arm. The involved gestures can be divided into two categories, according to whether human-object interaction happens:

- **Category 1:** the human gestures without human-object interaction;
- **Category 2:** the human gestures with human-object interaction.

Category 1 contains 7 upper body gestures: “be confident”, “have question”, “object”, “praise”, “stop”, “succeed” and “weakly agree”. Some examples are shown in Fig. 3. We can observe that, they are all natural gestures of intuitive meanings that are able to

Figure 3: The *Category 1* body gestures.

Figure 4: The *Category 2* body gestures.

reflect the human’s emotion state or behavior intention, not the ad hoc ones for specific applications. Therefore, gesture-to-meaning mapping is not needed in our system. Recognizing the gesture type can achieve the goal of gesture understanding directly. Because the humans’ behavior habits may be different, recognizing the natural gestures is much more challenging than the ad hoc ones but also more meaningful for the natural human-virtual human interaction. From the listed samples, it could be seen clearly that both human hand and upper body posture information are required to recognize these gestures. For instance, the upper body postures corresponding to “*have question*” and “*object*” are very similar. Without the hand gesture feature, they are difficult to distinguish. The same thing also happens to “*have question*”, “*weakly agree*” and “*stop*”. They correspond to similar hand gestures but very different upper body postures.

Category 2 possesses other 4 upper body gestures: “*call*”, “*drink*”, “*read*” and “*write*”. Fig. 4 exhibits the examples of the 4 gestures. Being different from the **Category 1** gestures, these 4 gestures are happening with human-object interactions. As discussed

in Sec. 3, such kinds of gestures are always ignored by the previous systems. One main reason is that objects may cause the human body occlusion, especially one the hand. In this case, vision-based hand gesture recognition methods will fail to work. However, this problem can be overcome by using CyberGlove. In this work, we attempt to recognize such gestures to make the human-virtual human interaction more lifelike. The same as the **Category 1** gestures, they are also recognized according to the human hand and upper body posture information.

As discussed above, how to describe the human hand and upper body posture is the key to recognize and understand the upper body gestures in the GUVHI system.

4.2 Feature Extraction and Fusion

In this subsection, we will introduce the feature extraction methods for human hand and upper body posture description. The way to fuse the multimodal features is also illustrated.

Figure 5: CyberGlove II data joints [Immersion 2010].

4.2.1 Hand Posture Feature

The Immersion wireless CyberGlove II is employed as the hand posture capture device in the GUVHI system. As one of the most sophisticated and accurate data gloves, the CyberGlove II can provide 22 high-accuracy joint-angle measurements in real-time. These measurements can reflect the bending degree of fingers and wrist. The 22 data joints (marked as big color dots) are located on the CyberGlove as shown in Fig. 5. We find that the wrist posture is not a stable feature for recognizing the users' hand gestures in our application. For the same hand gesture, the wrist bending degree of different users may change a lot. This phenomenon is caused by the users' different behaviour habits and hand sizes. So, the two wrist data joints (marked as red) will be discarded. A 20-dimensional feature vector F_{hand} will be extracted from the 20 yellow data joints to describe the human hand posture as

$$F_{hand} = (h_1, h_2, h_3 \cdots h_{19}, h_{20}), \quad (1)$$

where h_i is the bending degree corresponding to the yellow data joint i .

4.2.2 Upper Body Posture Feature

With the Kinect sensor, we shape the human upper body posture intermediately by the 3D skeletal joint positions. For a full human subject, 20 body joint positions can be tracked by the real-time skeleton tracker [Shotton et al. 2011] based on the Kinect depth frame, which is invariant to posture, body shape, clothing, etc. Each joint J_i is represented by 3 coordinates at the frame t as

$$J_i = (x_i(t), y_i(t), z_i(t)). \quad (2)$$

However, not all the 20 joints are necessary for the upper body gesture recognition. As aforementioned, head and right arm are highly correlated with the 11 upper body gestures needed to be understood. For simplicity, we only choose 4 descriptive upper body joints: "head", "right shoulder", "right elbow" and "right hand" shown as the big green dots in Fig. 6. These selected joints match with the right hand CyberGlove II.

The original 3D joint positions are sensitive to the relative position between the user and the Kinect. Directly using them is not stable for representing the body posture. Meanwhile, solving this problem by restricting the user's position is unfeasible for the practical application. In [Wang et al. 2012b], human action is recognized by making use of the pairwise relative positions of the joints. Inspired by this work, we choose the joint "middle of the two shoulders" marked as the red dot in Fig. 6 as a reference joint. And the pairwise relative positions between the 4 previously selected joints and

Figure 6: The selected body skeletal joints.

the reference joint are computed as

$$J_{sr} = J_s - J_r, \quad (3)$$

where J_s is one of the 4 selected joints and J_r is the reference joint. Obviously, J_{sr} is less sensitive to user-Kinect relative position. It is mainly determined by the body posture. The reason to choose "middle of the two shoulders" as the reference joint is that it can be robustly detected and tracked to most cases. And it is rarely blocked by the limbs or the objects in daily gestures. Finally, a feature vector F_{body} of 12 dimensions is constructed by combining the 4 pairwise relative positions to shape the upper body posture as

$$F_{body} = (J_{1r}, J_{2r}, J_{3r}, J_{4r}), \quad (4)$$

where J_{1r} , J_{2r} , J_{3r} and J_{4r} are the 4 selected pairwise relative positions.

4.2.3 Feature Fusion

Till now, two multimodal feature vectors: F_{hand} and F_{body} are extracted to describe the hand posture and the upper body posture respectively. To fully understand the human upper body gesture, the joint information of the two feature vectors is needed. Both of them are important for the recognition task. However, they are extracted from the different sensors and their values locate in very different value ranges. Simply combining the two feature vectors as the input for classifier will yield the performance bias on the feature vector of low values. Here, we propose to rescale the two feature vectors into the similar range before the feature fusion. Supposing F_i is one dimension of F_{hand} or F_{body} , F_i^{max} and F_i^{min} are the corresponding maximum and minimum value in the training set. Then F_i can be normalized as

$$\hat{F}_i = \frac{F_i - F_i^{min}}{F_i^{max} - F_i^{min}}, \quad (5)$$

for both training and test.

After normalization, the effect of the two feature vectors in classification will be balanced. And they are fused to form the integral feature vector for classification by concatenation as

$$\vec{F} = (\hat{F}_{hand}, \hat{F}_{body}). \quad (6)$$

Finally, a 32-dimensional feature $\vec{F}$ is constructed for upper body gesture recognition.

Figure 7: Illustration of the LMNN distance metric learning.

4.3 Classification Method

Using $\vec{F}$ as the input feature, the upper body gestures will be recognized by template matching based on the *energy-based LMNN classifier* proposed in [Weinberger and Saul 2009]¹. It is derived from the energy-based model [Chopra et al. 2005] and the LMNN distance metric learning method [Weinberger and Saul 2009]. The latter part is the key to constructing this classifier. LMNN distance metric learning approach is proposed to seek the best distance measure for the k -nearest neighbor (KNN) classification rule [Cover and Hart 1967]. As one of the oldest methods for pattern recognition, the KNN classifier is very simple to implement and use. Nevertheless, it can still yield comparative results in certain domains such as object recognition and shape matching [Belongies et al. 2002]. And it also has been applied to action recognition [Müller and Röder 2006] recently.

The KNN rule classifies each testing sample by the majority label voting among its k -nearest training samples. Its performance crucially depends on how to compute the distances between different samples for the k nearest neighbors search. Euclidean distance is the most widely used distance measure. However, it ignores any statistical regularities that may be estimated from the training set. Ideally, the distance measure should be adjusted according to the specific task being solved. To achieve better classification performance, LMNN distance metric learning method is proposed to mine the best distance measure for the KNN classification.

Let $\{(\vec{x}_i, y_i)\}_{i=1}^n$ be a training set of n labeled samples with inputs $\vec{x}_i \in \mathbb{R}^d$ and class labels y_i . The main goal of LMNN distance metric learning is to learn a linear transformation $\mathbf{L} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that is used to compute the square sample distances as

$$\mathcal{D}(\vec{x}_i, \vec{x}_j) = \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2. \quad (7)$$

Using $\mathcal{D}(\vec{x}_i, \vec{x}_j)$ as the distance measure tends to optimize the KNN classification by making each input $\vec{x}_i$ have k nearest neighbors that share the same class label y_i to the greatest possibility. Fig. 7 gives an intuitive illustration on LMNN distance metric learning. Compared with Euclidean distance, LMNN distance tries to pull the nearest neighbors of class y_i closer to $\vec{x}_i$, meanwhile push the neighbors from different classes away. Under the assumption that the training set and the test set keep the similar feature distribution, LMNN distance metric learning can help to improve the KNN classification result.

The energy-based LMNN classifier makes use of both the $\mathcal{D}(\vec{x}_i, \vec{x}_j)$ distance measure and the loss function defined for LMNN distance metric learning. It constructs an energy-based criterion function, and the testing sample is assigned to the class which yields the minimum loss value. Because the related theory is sophisticated, we do not give the detailed definition on the energy-based

¹The source code is available at <http://www.cse.wustl.edu/~kilian/code/lmnn/lmnn.html>.

Figure 8: The examples of Chloe's body actions and facial expressions.

LMNN classifier here. The readers can turn to [Weinberger and Saul 2009] for reference.

5 Human-Virtual Human Interaction

In the GUVHI system, as a case study, human and virtual human interact in the context of they are meeting to discuss something. The virtual human is a female named "Chloe". Based on the output of the human gesture understanding module, she can perceive and understand the human upper body gestures described in Sec. 4.1 and give response to human. For human-virtual human interaction, how to maintain the users' interest is important. In our implementation, Chloe is human-like to be capable of executing the reaction combining body action, facial expression and verbal language. The three behaviour factors can reflect the virtual human's personality and emotional state. Adding them together to Chloe can make her more conceivable and believable. They will provide the users with more vivid feedback. Fig. 8 exhibits some examples of Chloe's body actions along with the corresponding facial expressions. As shown, she is under the emotional states of "happy", "moderate" and "sad" respectively, which can be clearly distinguished through the visual difference. Nonverbal behaviors can help to structure the processing of verbal information [Krämer et al. 2003]. Thus, body action and facial expression can enhance the performance of Chloe's verbal language to make the users more impressed.

As the real human, virtual humans' behaviors are associated with their perception. At current stage, Chloe's reaction is mainly triggered the human body language. Here, we propose a scenario to control Chloe's behaviors from body action to verbal language. This scenario covers all the 11 upper body gestures. And it is pri-

Table 1: The scenario for human and virtual human interaction.

Human body gestures	Chloe's reaction	
	Facial expression	Verbal language
"be confident"	happy	Great to see you so confident.
"have question"	moderate	What is your question?
"object"	sad	Why do you disagree with me?
"praise"	happy	Thank you for your praise.
"stop"	moderate	Why do you stop me?
"succeed"	happy	Congratulations on your success.
"weakly agree"	happy	OK, lastly we reach agreement.
"call"	moderate	You can make your call first. I can wait.
"drink"	moderate	Do you need more drink?
"read"	moderate	You can read it slowly. Do not worry.
"write"	moderate	If you need time for record, I can slow my talk.

marily designed according to the real human's reactions to these gestures. It aims to simulate the natural human-human interaction, even when some accidents happen, such as one human has to answer a coming call suddenly. By this way, the human-virtual human interaction will be more lifelike. Because it is difficult to describe the virtual human's body actions accurately in text, only Chloe's facial expression and verbal language in the scenario are illustrated in Table 1. Chloe's facial expressions are limited to "happy", "moderate" and "sad" which are highly correlated with her emotional states corresponding to the gestures. And the verbal language is relatively simple to make the users understand it easily.

6 Experiment and Discussion

To verify the effectiveness of the proposed upper body gesture understanding method, a human upper body gesture dataset is constructed by us. This dataset involves all the 11 upper body gestures being recognized. The gesture samples are captured from 23 people of different genders, body sizes and races. During the sample collection, no strict constraint was applied to the persons. They did the gestures by their own habits. The user-Kinect relative position was also not strictly limited. And the CyberGlove II was initially calibrated only once for all the people using a standard calibration. That was a convenient way for the users. According to the dataset construction configuration above, gesture diversities may exist among the people. This causes the difficulties for gesture recognition. However, our dataset is built up closely to the practical use. Fig. 9 exhibits parts of the **Category 1** and the **Category 2** gesture samples ("have question", "succeed", "call" and "drink") captured from 6 people for comparison. For brevity, not all kinds of the gestures are shown. The 5 body skeletal joints proposed in Fig. 6 are marked as the color dots in the sample images according to the output of the skeleton tracker [Shotton et al. 2011] employed by us. And they are connected by the straight segments to provide an intuitive way to shape the upper body posture. From the exhibited samples, it could be observed that:

- The listed gestures can be mainly distinguished based on the hand posture and the upper body posture for the different people. However, the people do the same gesture differently to some degree. This phenomenon actually leads to the difficulties for gesture recognition;
- For the different people and the different gestures, the 5 body skeletal joints used for gesture recognition can be tracked robustly, even when the human-object interaction happens. And their resulting positions are accurate enough for the specific application. Meanwhile, the CyberGlove II is a human-touch device that can capture the hand posture robustly to yield high-

Table 2: Classification result (%) of the constructed dataset. The best performance is shown in boldface. Standard deviations are in parentheses.

Classifiers	Training sample number per class		
	4	6	8
KNN (Euclidean)	87.75(± 3.78)	90.59(± 2.58)	92.00(± 1.17)
KNN (PCA)	75.22(± 6.46)	81.71(± 5.08)	85.82(± 4.55)
KNN (LDA)	52.15(± 12.08)	70.37(± 10.90)	69.58(± 18.97)
KNN (LMNN)	88.90(± 3.30)	90.80(± 2.99)	92.73(± 1.77)
Energy (LMNN)	90.62 (± 0.99)	94.55 (± 2.81)	94.91 (± 1.01)
Classifiers	10	12	14
KNN (Euclidean)	92.59(± 2.07)	92.73(± 3.06)	93.54(± 1.97)
KNN (PCA)	88.39(± 2.78)	87.27(± 4.91)	86.06(± 5.60)
KNN (LDA)	87.83(± 8.69)	90.41(± 8.48)	90.30(± 6.57)
KNN (LMNN)	93.15(± 2.00)	92.56(± 2.80)	94.55(± 1.53)
Energy (LMNN)	96.36 (± 1.35)	97.02 (± 1.71)	97.78 (± 1.94)

accuracy data. Therefore, the GUVHI system can acquire available data for gesture recognition stably.

Currently, the upper body gestures in our system are almost static ones. For a person, we will pick up one representative frame for each gesture to construct the dataset. Thus, the resulting dataset contains $23 \times 11 = 253$ gesture samples in all. In experiment, the available dataset samples are randomly split into a training set and a testing set for 5 times, and the average accuracy is reported.

The KNN classifier is employed as the baseline to make comparison with the energy-based LMNN classifier. They will be compared both on classification accuracy and time consumption. The KNN classifier will run with different kinds of distance measures. Following [Weinberger and Saul 2009], " k " is set as 3 for all the cases. Additionally, the ball tree data structure [Beygelzimer et al. 2006] is applied to speeding up the k -nearest neighbors search for the two classifiers both in training and test. As demonstrated in [Beygelzimer et al. 2006], this data structure works well for the low dimensional gesture feature vector $\vec{F}$ defined by us. Because training sample number is a crucial factor for accuracy, the two classifiers will be compared corresponding to different amounts of training samples. The training sample number ranges from 4 to 14 for each class with the step size 2.

Other two well known distance metric learning methods: PCA [Jolliffe 1986] and LDA [Fisher 1936] are used for comparison with the LMNN distance metric learning approach. For PCA, the first 10 eigenvectors are used to capture roughly 90% of the sum of eigenvalues. The distance measures yielded by PCA and LDA will be applied to the KNN classifier.

Table 2 lists the classification results of different classifiers. We can observe that:

- The upper body gestures in the dataset can be well recognized by the proposed gesture recognition method. More than 97.00% classification accuracy can be achieved if the enough training samples are used. With the increase of training sample amount, the performance is generally improved consistently;
- Corresponding to all the training sample numbers, the energy-based LMNN classifier can yield the highest classification accuracy. Even with small number (such as 4) of training samples, it can still achieve good performance (90.62%). When the training sample number reaches 14, the classification accuracy (97.78%) is relatively very high. And its standard deviations are the lowest for almost all the cases, which means that the energy-

Table 3: Average testing time consumption (ms) per sample. The program is running on the computer with Intel (R) Core (TM) i5-2430M @ 2.4GHz (only using one core).

Classifiers	Training sample number per class		
	4	6	8
KNN (LMNN)	0.013	0.017	0.018
Energy (LMNN)	0.075	0.088	0.104
	10	12	14
KNN (LMNN)	0.021	0.024	0.029
Energy (LMNN)	0.129	0.145	0.178

based LMNN classifier is also robust to the gesture diversities among people;

- KNN classifier can also yield good result on this dataset. However, it is inferior to the energy-based LMNN classifier. Compared to Euclidean distance, LMNN distance metric learning method can improve the performance of KNN classifier consistently to most cases. However, it works much better on the energy-based model;
- PCA and LDA do not work well on this dataset. Their results are even worse than the Euclidean distance. The main reason is that the two distance metric learning methods need large number of training samples. And that is their limitation for the practical application.

Besides the classification accuracy, another important factor that we concern about is the testing time consumption of our gesture recognition method. The reason is that the GUVHI system should be running in real time. The energy-based LMNN classifier and the LMNN KNN classifier are two classifiers with the highest accuracies. Here, we make a comparison on their time consumption. Table 3 lists the average running time per testing sample. It could be observed that both the two classifiers are extremely fast for the application. The time consumption mainly depends on the number of training samples. Actually, the LMNN KNN classifier is much faster than the energy-based LMNN classifier. If large number of training samples are used (such as tens of thousand), the LMNN KNN classifier will be the better choice.

7 Conclusion and Future Work

The natural human-virtual human interaction is an important component for many practical virtual systems. In most existing applications, human and virtual human communicate mainly based on verbal language. This paper proposes to use natural human body language for human-virtual human interaction. The GUVHI system, a novel human upper body gesture understanding and interaction system, is implemented. At the current stage, 11 human upper body gestures with and without human-object interaction can be perceived and understood by virtual human in this system. Meanwhile, the virtual human responses to human gestures with its own personality. It reacts by the means of body action, facial expression and verbal language simultaneously, aiming to give users the human-human interaction experience.

For the GUVHI system, a new human-gesture capture sensor solution is suggested, which is the most important contribution of this paper. In our system implementation, the CyberGlove II is the hand posture capture device, and the Kinect is employed to acquire the upper body posture. Based on the gesture data from the sensors, descriptive gesture feature is proposed. LMNN metric learning and the energy-based classifier are used for classification to enhance the gesture recognition performance. For the experiment, a human up-

per body gesture sample dataset is constructed. It contains human gesture samples from 23 people. The experiments verify the effectiveness of our gesture recognition algorithm both on accuracy and time consumption.

So far, the human gestures involved in the GUVHI system are static ones, e.g., “have question”, “praise”, “call” and “drink”, etc. In future work, we plan to make the virtual human understand dynamic gestures, such as “wave hand”, “type keyboard”, “shake hand” and “clap”, etc. Speech recognition can be added to make the system more stable and natural for the human-virtual human interaction.

Acknowledgements

This research is partially supported by IMI Seed Grant. It is partially carried out at BeingThere Centre with the support from the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

We will thank the people: Zhengpeng Bian, Yong Chen, Derek Ho, Lu Dong, Hui Fang, Guibing Guo, Junhui Hou, Jesslim, Khin Yaminnl, Bingbing Li, Hui Liang, May Thu Aungl, Nikhil, Qui, Jianfeng Ren, Rohit Kumar Dubey, Shantanu Choudhary, Yan Xia, Zerrin Yumak, Juzheng Zhang, Wenjing Zhang, Zhijun Zhang and Hongyuan Zhu in IMI @ NTU a lot. They helped us to establish the human upper body gesture dataset for experiment.

In addition, we need to especially thank Hui Liang, Flavien Picon and Shantanu Choudhary. They gave us much assist to implement the GUVHI system.

References

- ADAMO-VILLANI, N., HEISLER, J., AND ARNS, L. 2007. Two gesture recognition systems for immersive math education of the deaf. In *Proceedings of the First International Conference on Immersive Telecommunications*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 9.
- BELONGIES, S., MALIK, J., AND PUZICHA, J. 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 4, 509–522.
- BERMAN, S., AND STERN, H. 2012. Sensors for gesture recognition systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 42, 3, 277–290.
- BEYGELZIMER, A., KAKADE, S., AND LANGFORD, J. 2006. Cover trees for nearest neighbor. In *In Proceedings of the Twenty Third International Conference on Machine Learning (ICML 2006)*, 97–104.
- CHOPRA, S., HADSELL, R., AND LECUN, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1, 539–546.
- COVER, T., AND HART, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions in Information Theory* 13, 1, 21–27.
- FISHER, R. A. 1936. The use of multiple measures in taxonomic problems. *Annals of Eugenics* 7, 179–188.
- GOBRON, S., AHN, J., GARCIA, D., SILVESTRE, Q., THALMAN, D., AND BOULIC, R. 2012. An event-based architecture to manage virtual human non-verbal communication in 3d chatting

have
question

succeed

call

drink

Figure 9: The gesture samples captured from different people.

environment. In *Articulated Motion and Deformable Objects*. Springer, 58–68.

GRATCH, J., WANG, N., OKHMATOVSKAIA, A., LAMOTHE, F., MORALES, M., WERF, R. J. V. D., AND MORENCY, L. P. 2007. Can virtual humans be more engaging than real ones? In *Human-Computer Interaction, HCI Intelligent Multimodal Interaction Environments, LNCS*, vol. 4552, 286–297.

IMMERSION. 2010. Cyberglove II specifications. Available at <http://www.cyberglovesystems.com/products/cyberglove-ii/specifications>.

JOLLIFFE, I. T. 1986. *Principal Component Analysis*. Springer-Verlag.

KRÄMER, N. C., TIETZ, B., AND BENTE, G. 2003. Effects of embodied interface agents and their gestural activity. In *Intelligent Virtual Agents*, Springer, 292–300.

LU, G., SHARK, L.-K., HALL, G., AND ZESHAN, U. 2012. Immersive manipulation of virtual objects through glove-based

hand gesture interaction. *Virtual Reality* 16, 3, 243–252.

MÜLLER, M., AND RÖDER, T. 2006. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, Eurographics Association, 137–146.

PIPLICA, A., DELEON, C., AND MAGERKO, B. 2012. Full-body gesture interaction with improvisational narrative agents. In *Intelligent Virtual Agents*, Springer, 514–516.

RAUTARAY, S. S., AND AGRAWAL, A. 2011. Interaction with virtual game through hand gesture recognition. In *Multimedia, Signal Processing and Communication Technologies (IMPACT), 2011 International Conference on*, IEEE, 244–247.

RIZZO, A. A., LANGE, B., BUCKWALTER, J., FORBELL, E., KIM, J., SAGAE, K., WILLIAMS, J., ROTHBAUM, B. O., D-IFEDE, J., REGER, G., ET AL. 2011. An intelligent virtual human system for providing healthcare information and support. Tech. rep., DTIC Document.

- SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., AND BLAKE, A. 2011. Real-time human pose recognition in parts from single depth images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 1297–1304.
- STEDMON, A. W., PATEL, H., SHARPLES, S. C., AND WILSON, J. R. 2011. Developing speech input for virtual reality applications: A reality based interaction approach. *International journal of human-computer studies* 69, 1, 3–8.
- SWARTOUT, W., TRAUM, D., ARTSTEIN, R., NOREN, D., DEBEVEC, P., BRONNENKANT, K., WILLIAMS, J., LEUSKI, A., NARAYANAN, S., PIEPOL, D., ET AL. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents*, Springer, 286–300.
- TELEB, H., AND CHANG, G. 2012. Data glove integration with 3d virtual environments. In *Systems and Informatics (ICSAI), 2012 International Conference on*, IEEE, 107–112.
- WANG, C., MA, Q., ZHU, D., CHEN, H., AND YANG, Z. 2012. A methodology of real-time 3d virtual human motion control based on depth-sensing camera for agricultural machinery training. *Mathematical and Computer Modelling*.
- WANG, J., LIU, Z., WU, Y., AND YUAN, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 1290–1297.
- WEINBERGER, K. Q., AND SAUL, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research* 10, 207–244.
- ZHAO, W., XIE, X., AND YANG, X. 2012. Control virtual human with speech recognition and gesture recognition technology. In *Advances in Electrical Engineering and Automation*. Springer, 441–446.