An Immersive Kendo (Gum-do) Game with An Intelligent Cyber-Fighter

Jungwon Yoon[a], Se-Hwan Kim[b], Jeha Ryu[a] and Woontack Woo[b]

[a]Department of Mechatronics
[b]Department of Information and Communications
Kwangju Institute of Science and Technology (K-JIST)
1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
garden@geguri.kjist.ac.kr

ABSTRACT

This paper presents a new framework of an immersive kendo game with an intelligent cyber-fighter, which has its own internal needs, motivations, sets of multimodal sensors, a motor system, and a behavior system. Unlike conventional interface such as keyboard or joystick, the proposed system provides more natural and comfortable interface by exploiting multimodal interfaces such as 3D vision and speech recognition. In addition, the proposed 3D vision-based interface allows relatively free-movement in 3D space, when it compares with wired tracker-based interfaces. As a result, the user with real sword can experience an immersive fighting with the cyber-fighter in virtual environment. The proposed framework will have wide variety of applications in VR-based edutainment applications.

Keywords: Kendo game, multimodal interface, 3D vision, speech recognition, intelligent cyber-fighter

1. INTRODUCTION

Virtual Reality (VR) technology is flourishing with the rapid development of computer and related technologies. There are wide-range of VR applications such as entertainment, training, education, engineering, medical operation, teleoperation, and so on. Among them, the entertainment applications are the most popular and have a large commercial market.

A number of researchers have studied on building systems in which animated agents interact with a user. Bate's Woggle world combines behavior models with ideas from traditional animation to explore what he calls "believable agents"[1]. Even though Bate's agents are able to have fairly complex interactions, the mouse is the only way to interact with the user. Fisher's Menagerie[2] system also allows a user to interact with animated agents in real time using goggles. Yoon et al.[3] have implemented a character that can be trained using "clicker training" technique named Sydney K9.0. They included a module named DogEar that is designed for collecting real-world acoustic data integrated into the creature kernel's perception system. However, these systems tend to distract users by requiring conventional interfaces such as keyboard, mouse, gloves, goggles, and/or helmet. These interfaces have limitations in providing immersive interactions because they have to be worn or attached on the body and connected to computers with wires.

Meanwhile, ALIVE[4] system applies more complex combination of both behavioral and motivational complexity in the creature model, and use a vision interface with which these behaviors can use the user's actual position, body pose, and hand gestures as sensory input. However, the ALIVE system has a limitation in providing natural interaction since system mainly exploits visual input.

In this paper, we propose a framework for VR kendo simulation game. In order to provide an immersive experience for the user the VR system need to interact with a user with an intelligent agent. The proposed system consists of three modules: (i) a comfortable 3D vision-based interface, (ii) artificial intelligence (AI) with a multimodal interface (vision and speech), and (iii) immersive feedback by sound and screen. First, the vision-based 3D interface allows the user to move around freely without wires, as well as to exploit 3D information about a user and some objects in his/her hand. Second, an intelligent fighter, a virtual agent with AI, provides intelligent interaction by vision and speech recognition. Finally, big screen with sound feedbacks help user experience an immersive fighting experience.

This paper is organized as follows. In chapter 2, we describe about the system configuration of the proposed kendo game system. A multimodal interface, consisting of vision-based 3D interface and speech-based interface, is explained in chapter 3. Chapter 4 and 5 describe the intelligent agent and virtual environment, respectively. Finally, experimental results and discussions are followed in chapter 6 and 7, respectively.

2. DESCRIPTION OF THE KENDO GAME SYSTEM

A kendo is one of fencing sports with a bamboo sword and light protective armor. Fighters wear protective

equipments covering target areas: head, wrists, and abdomen. To make a valid cut a player must strike his opponent with the bamboo sword on target areas. The points are scored by striking any part of target areas on the opponent. Fighters are required to shout out the name of the target while they are striking (or thrusting). The first contender who strikes target areas three times becomes a winner.

In order to properly simulate the physical kendo sports in a VR setting, visual and auditory feedback should display collision situation in real time for better feeling when two swords are collided each other. Moreover, a wide motion range should be provided by a vision interface for better virtual reality. Fig. 1 shows the proposed kendo game system with vision and speech interfaces. When two swords collide with each other or one sword strike a target area in a body, collision situations are determined and are reflected by the sound to the human swordsman. The big screen displays virtual environment with swordsman and fighting floor. The first person viewpoint is selected so that an operator may see the opponent in virtual environment.

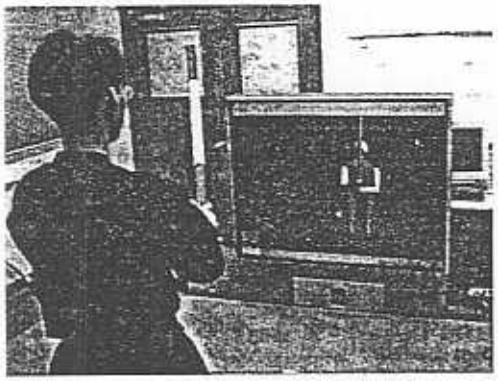


Figure 1. An immersive kendo game with audio-visual feedback.

Main feature of the proposed kendo game is that both motion of a bamboo sword grasped by the human hand and motion of the human body are acquired by a vision interface with a multiview camera. Two colored markers (red and green) are attached to both ends of the bamboo sword in order to compute the pose of the bamboo sword by color segmentation of the foreground object, which will be found with background segmentation. The center position of the human body is detected by a multiview camera.

Next, the three words of "Meo-Ri (HEAD)", "Heo-Ri (ABDOMEN)" and "Son-Mok (WRIST)" from a fighter shouting the target points are recognized by simple speech recognition. This speech recognition will affect an opponent action in a virtual environment which has its own needs and motivation. When a user pronounces "Meo-Ri (HEAD)", "Heo-Ri (ABDOMEN)" and "Son-Mok (WRIST)", unique features of those words are saved. After that, new speech inputs are compared with the saved features and the classified words are finally transferred to the virtual fighter. The striking and the shouting target area should be the same since a fighter should shout the target point while striking target area to get one score and should be detected simultaneously. However, it is difficult to detect simultaneously two situations of collision detection and speech recognition in reality due to the difference of real time motion and irregular duration of human speech. Therefore, we will use speech recognition only to give additional information at opponent action generated by computer.

In a one-fighter mode, a human fighter is fighting with the computer-generated cyber opponent who has artificial intelligence. In this mode, a virtual environment will generate intelligent motion of the virtual fighter in response to the real human fighter motion. To give intelligence to a virtual fighter, the behavior, motivation, motor, and perception system are defined in virtual fighter. Next, we will get the final actions through the kernel operation of the virtual fighter.

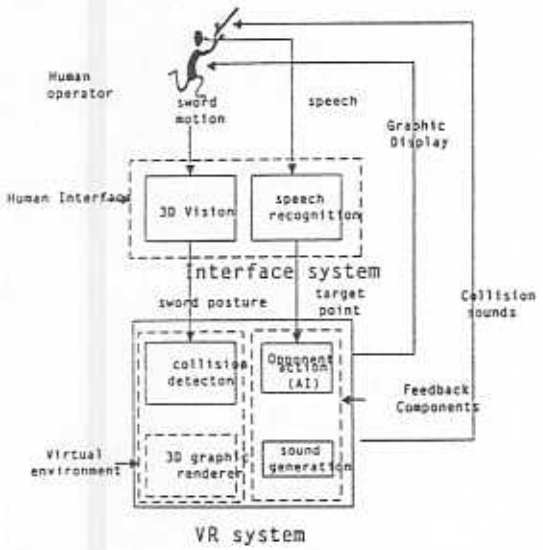


Figure 2. Kendo Game System Block Diagram

Fig. 2 shows system block diagram of the proposed kendo game system. The position of a sword and the speech recognition of a user are communicated from one

PC to another in real time. The proposed system of a kendo game is composed of an interface system and a VR system. The interface system computes 3D information from human body and sword and executes speech recognition. The VR system performs graphic simulation, collision detection[5] after receiving sword positions from vision interface, and AI. It also displays virtual environment and transfers collision sounds to an operator. For interface of user input with computer, the motion information is acquired with vision interface taken by multiview images and transfers the position information of bamboo sword and human body into the virtual environment and artificial intelligence part. The interface system and VR system is connected via the TCP/IP protocol networking by using the win socket class in MFC of Visual C++.

3. MULTIMODAL INTERFACE

3.1. Vision-based 3D Interface

Although sensor or marker-based 3D motion tracking provides relatively accurate position and movement information over vision-based tracking, they have several inherent problems. The devices need complicated calibration and conversion procedures to get accurate movement data. Given accurate position and movement data, there still remain several other hindrances obstructing widespread usage of 3D-interface[6]. Therefore, we selected the vision interface to find information in 3D space.

In order to get 3D information from multiview images, we first segment the user from the background and estimate depth information using multiview images. Given depth information we estimate the center of the segmented body. Next, we track the movement of the body by observing the center of the segmented body. Finally, we segment two colored markers located at the end points of sword from the foreground image and estimate the center of the segmented two colored markers to find the pose of a bamboo sword.

Obviously, unlike such complicated motion data acquisition equipment, such as heavy headsets, data gloves and tethers, which attaches infrared or magnetic sensors to the user, the proposed vision-based 3D interface does not distract the user since it tracks the movement based on invisible center of objects exploiting depth information estimated using multiview images.

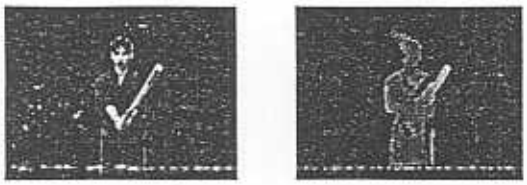
3.1.1. Segmentation

The first step is to segment the user from a natural background scene to get the information about an object. Even though the user can be separated from a background using blue screen technique, in this paper we do not consider such a special environment. Instead, the used scheme segments the user from a natural scene by exploiting multiple cues such as intensity (or color), edge, motion and depth [7,8].

We use a robust segmentation scheme jointly exploiting color (or intensity), edge, motion and disparity information[8,9]. The proposed moving object segmentation scheme consists of three steps, which are (1) static background scene capture and estimation of its statistics (2) disparity estimation (3) moving object segmentation from a natural scene. The basic idea of separating moving objects from a background is to subtract a current image from the reference images which are acquired from a static or non-static background during a period of time. To alleviate the effect of lighting condition we adopt a normalized (r,g,b) color space with statistics of the reference images. The reference background image and statistical parameters are computed over a number of frames without objects. The background is modeled statistically on pixel by pixel basis, using the mean and variance of each normalized color component.

We first capture a static background scene and estimate corresponding statistics for each pixel in the image, N(m; ¢), where m and ¢ denote mean and standard deviation, respectively. The statistics N (m; ¢) are used in order to decide threshold values in the initial segmentation process. Next, we estimate a smooth disparity of current scenes. Finally, we segment moving objects from the static background scene by comparing intensity (color), edge, motion and disparity. To segment only objects of interest we assume that these objects have a limited range of disparities, i.e. depth, and that the disparities of the objects change smoothly. The segmented object containing depth information is ready to be used in object tracking. According to the experimental results[8], the proposed hybrid segmentation scheme efficiently separates the user not only from blue screen but also from a real scene.

Fig. 3 shows the results of segmentation. Figure 3(a) is the segmented image after separating foreground from background using statistical information. Figure 3(b) is its disparity image.



(a) (b)

Figure 3. User Segmentation

3.1.2. Human and Sword Tracking

After user segmentation, the position information of bamboo sword and human body can be estimated using the centroid of each object. To get 3D information of bamboo sword from foreground image, we segmented colored marker images from user images which are segmented from background. We used the special-colored (red and green) markers at the end of bamboo sword so that we may easily detect those objects. If we set the matching point as a centroid point of colored marker at the one end of bamboo sword, we can estimate the position of bamboo sword in 3D space based on two 3D centroid points which are found by color segmentations as shown in Fig. 4.

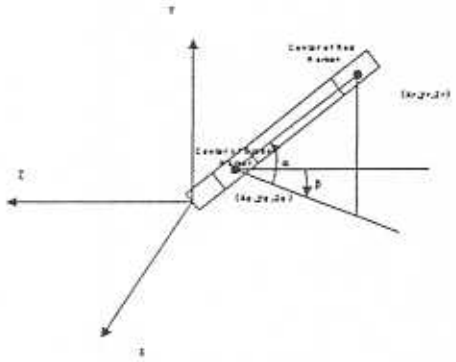


Figure 4. Colored Marker at Bamboo Sword

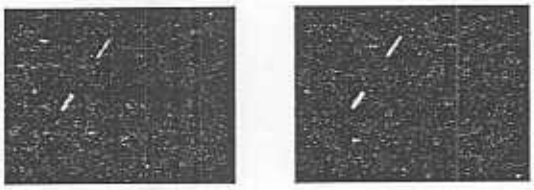
The azimuth angle α and elevation angle β of bamboo sword can be calculated as (see Fig. 4)

$$L_1 = \sqrt{(x_r - x_g)^2 + (y_r - y_g)^2 + (z_r - z_g)^2} \quad (1)$$

$$\alpha = \tan^{-1}((x_r - x_g)/(z_r - z_g)) \quad (2)$$

$$\beta = \sin^{-1}((y_r - y_g)/L_1) \quad (3)$$

Next, we can find the position of human object in the 3D coordinate by estimating center of user images which are subtracted colored marker images from user images. Finally, each end position of bamboo sword and the position of human object will be transferred to the virtual fighter kernel, combined with the results of speech recognition. Fig. 5 shows the results of color segmentation. Fig. 5(a) is the segmented image based on the information of color labels of bamboo sword after separating foreground from background. Fig. 5(b) is its disparity image.



(a) (b)

Figure 5. Color Segmentation

3.2. Speech-based Interface

Speech inputs for the proposed system are simply classified into three words only: "Meo-Ri (HEAD)", "Heo-Ri (ABDOMEN)" and "Son-Mok (WRIST)". Our speech recognition system is mainly developed for clearly differentiating the above three different words in a simplest possible way in order to have minimum time-delay in the speech recognition processing for real-time interaction. Feature extraction of speech is therefore adequate for our kendo system since we only need to differentiate three simple words. We applied the spectrogram for feature extraction. Fig. 6 shows the resultant features of words from several experiments and it shows clear differences among three words, which makes it easy to differentiate the words by feature extraction.

There are some difficulties in speech segmentation by the spectrogram since the speech length is always different and calculating the difference between pre-defined pattern and a new input speech is difficult and speech sound has always different intensity. The intensity of start portion of the speech is always different and unstable. Therefore, we try to find the start position of speech by threshold and get the signal within specific time window. If users are trained by experiment, the speech recognition success rate is over 80%. Even though speech recognition is not perfectly correct, users can be more tolerant of imperfections in a virtual fighter's

perception since a human fighter cannot expect reactions of the virtual fighter.

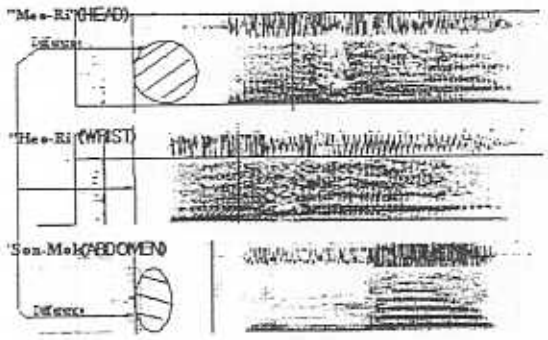


Figure 6. Spectrogram Representation

4. INTELLIGENT AGENT

Our virtual fighter has a motivation system, a perception system to perceive their environment, a behavior system which they can perform, a motor system which chooses the set of activities to perform given the internal needs of the agent and the opportunities presented by the environment. The fighter's state and geometry are updated according to the motor activities with the chosen behavior and rendered on every time-step. The position and velocity of human user's location, sword direction, auditory input affect the behavior of the virtual fighter, and the user receives visual as well as auditory feedback about the virtual fighter's internal state and reactions.

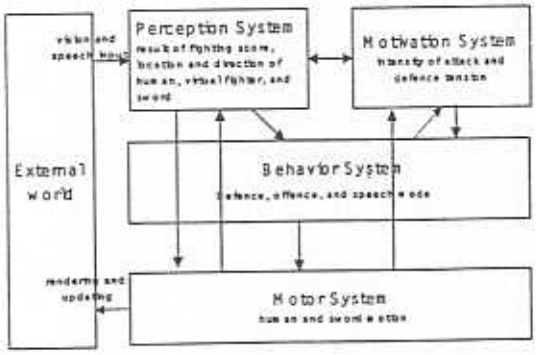


Figure 7. Schematic Diagram of an Opponent Fighter Action

Fig. 7 shows a schematic diagram of the opponent virtual fighter action. Arrows represent information flows between components. The motivation system is composed of state and drive. State refers to the fighter's tension and is associated with the fighter's action. Drives are tendency to fight in given game. The behavior system performs relevant action given the perception and motivation input. The behavior system indicates the primitive action to motor system which is composed of primitive skills and motions. The perception system has the information inside and outside of the fighter and the motivation system is provided as the means for enabling virtual fighter to perform certain skills and display them on the screen.

4.1. Perception (Movement and Speech)

Kendo fighter's perception system can cope with real time voice inputs from a human fighter through speech recognition. The distance difference of virtual opponent fighter in VR and human fighter will affect the state of virtual fighter. The direction of the sword of a human fighter in VR will also give the right and left information of virtual opponent fighter's action selection of sword. The velocity of opponent fighter and human fighter in VR can be recognized in high, middle, and low speed action. This movement of a human fighter is one input and anther one is speech. In this system, we mainly consider three words, this is, "Meo-Ri (HEAD)", "Heo-Ri (ABDOMEN)" and "Son-Mok (WRIST)". Since we only have to differentiate which word is pronounced by the user, we can maintain the simplicity of the system. In order to extract feature of voice, we applied the spectrogram and classified three words.

4.2. Reaction

4.2.1. Motor system

The motor system does real time motion interpolation to produce motion with the appropriate art contents. We can classify the primitive actions of fighter as human body motion and sword motion in motor system. Body motion can be classified as stop, move forward, move backward, move right, move left, and so on. Sword motion can also be classified such as stop, stop with stand, striking head, right and left abdomen, and striking right and left slope.

4.2.2. Motivation system

The motivation system consists of two parts: the drive system and the state. The drive system is divided into attack intensity and defense intensity varying over the score of fighting. The state can be divided into five categories with perception system. Each state can be

categorized as; high tension, tension, average, relaxed, and very relaxed.

4.2.3. Behavior system

Behavior action selection is done according to the motivation and perception systems. The behavior system has defense mode, offense mode, and speech mode. This is the tendency of a virtual opponent fighter and action will be changed with the motivation system and perception system. If a speech input enters into behavior system, priority of a speech mode becomes high in a behavior system. All the behavior system is composed of hierarchical structure and each motion should have smooth action in a given condition. Therefore, all possible sword motions are classified hierarchically. Each hierarchical network is selected according to the perception system or motivation system.

5. VIRTUAL ENVIRONMENT

For a virtual environment, a graphic environment has been developed by using OpenGL API. The human geometric model is composed of a spherical head, a cylindrical body and cylindrical upper and lower arms. The first person viewpoint is selected so that an operator may see the opponent in virtual environment. The virtual environment with an opponent human model and the floor changes according to the motion change of the manipulated human model so that the operator may feel realistic and intuitive. In the proposed kendo game system, collisions may occur between two swords and between a sword and one of the target areas: head, wrist, and abdomen.

For sound feedback, we use the direct sound-x classes. By this class, we record sounds such as the sounds of collisions between sword and sword and shouting voices such as "Meo-Ri (HEAD)", "Heo-Ri (ABDOMEN)" and "Son-Mok (WRIST)" which will be used during fighting for an opponent swordsman controlled by computer. Fig. 8 shows the whole virtual environment procedure of the proposed kendo game. Virtual environments with update rate of 20Hz are composed of collision detection, graphic rendering, AI, feedback graphic display, collision sound to operator and speech recognition. Through vision and speech interface, the positions of human and sword are transferred to VR environment. This information is used to update the graphic rendering and collision detection and make the opponent action according to the kernel operation of virtual fighter.

5.1. Procedures of Virtual Environment

As shown in Fig. 8, the virtual environment is processed in the following steps:

i) Obtain the positions of a real sword and human body by vision interface.

ii) Transform from the real position to virtual environment world position.

iii) Perform collision detection and find intention of swordsman

iv) Determine the reactions of opponent swordsman in virtual environment by AI procedure.

v) Display the reaction of swordsman in screen and feedback sound of collision.

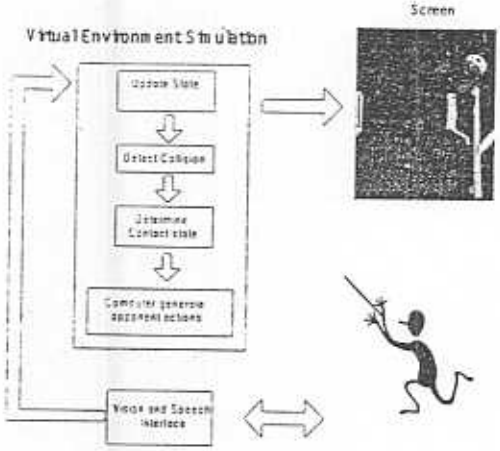


Figure 8. Procedures of Feedback Actions

5.2. Motion Mapping between Human/Sword Model and Two Motion Objects (Real Human Body and Sword)

The motions of the human model and sword are controlled by a vision interface. The x and y coordinate of the human body center manipulate the x and y translational motions of whole human model and sword while the z-rotation of human body makes simultaneous rotation of the z-axis of the whole human model and sword. The rotations about the x and z axes of bamboo sword calculated by two colored marker centers cause the rotations of the 2-dof pitch and yaw motions of the bamboo sword. The pitching motion ϕ of a sword makes the subsequent rotations of shoulder, elbow, and wrist of a virtual fighter arm model shown in Fig. 9. These rotations are related to the sword pitching angle such that $\phi = c_1\alpha_1 = c_2\alpha_2 = c_3\alpha_3$ where c_1, c_2, c_2 are empirically predefined constants for realistically looking arm motion. However, the yawing motion θ of the sword is not affecting the motions of human model.

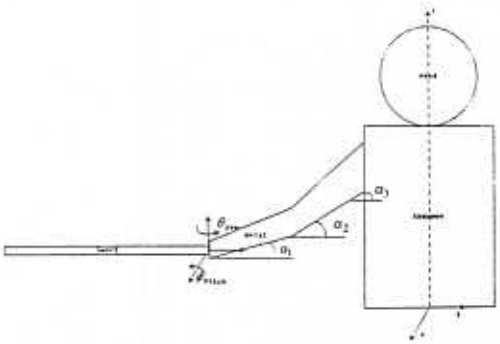


Figure 9. Motion Mapping between Virtual Human /Sword and Real Bamboo Sword Motions

6. SIMULATION RESULTS

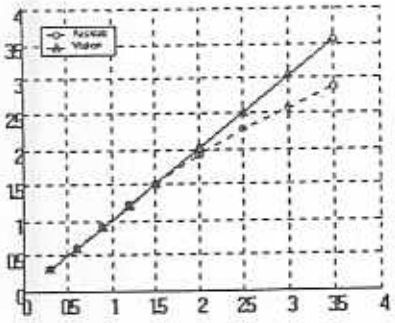
We set an adequate disparity range to enhance computation speed and get an available resolution. The speed of objects calculated by vision interface is limited to 5 Hz due to the time consumption of 3D disparity calculation of multiview images. Through the observation of the experiment, we found that the tracking performance of a human is sufficient to enjoy a real fighting with a virtual swordsman in virtual environment and tracking is stable with a rapid human motion. Table 1 shows specifications of the kendo system by experiments. As shown in Fig. 10, vision-based system is more adequate for the kendo system than tracker-based system. The vision-based kendo game system has more comfortable interface, while maintaining tracking accuracy and speed, when it compares with the performance of tracker-based system (POLHEMUS FASTRAK).

Table 1. Specifications of the kendo system

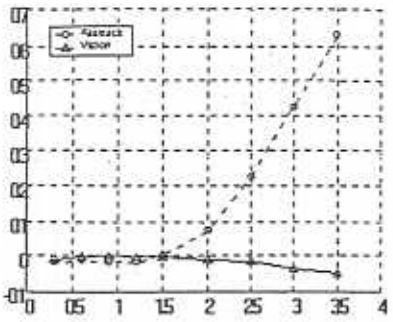
Parameters	Measurements
Fighting Area	200cm (w) * 400cm (h)
Sword Resolution	5 deg (pitch), 5 deg (roll)
Body Resolution	0.05m
Bandwidth of System	5Hz

Fig. 10(a) shows the measured distance depending on the real distance from a camera to a real object and from a FASTRAK receiver to a transmitter, respectively. In Fig. 10(b), the measurement error according to the distance is shown. According to the results, the performance of vision-based system is better than that of tracker-based system. For example, when the distance is small or within a proper range (about 2.0 m), we cannot differentiate the difference in performance between them. However, as the distance increases the performance difference becomes apparent. We can see that the measured distance is much more accurate when we use the vision-based system. As described in the FASTRAK specifications, FASTRAK provides the specified accuracy only when standard receivers are located within 76 cm of the standard transmitter even though an operation with separations up to 305 cm is possible with reduced accuracy[10]. Thus, the specifications prove the experimental results are correct. As shown in Fig. 10(c), even standard deviation becomes much deteriorated as the distance increases in case of tracker-based system compared to vision-based system. From Fig. 10(c), we observed that there is a little strange value when the distance is 0.3 m compared with other values. The reason is some restriction by disparity range. This is the reason that we have an abrupt value at that distance.

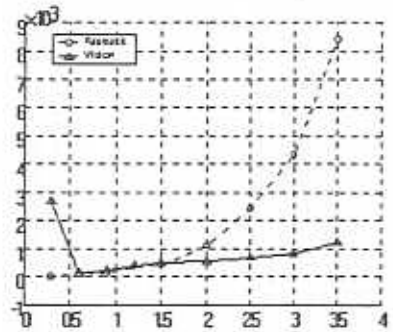
In conclusion, in most of the VR systems, the tracker-based system is not suitable compared with the vision-based system. Even though some products including ISOTRAK, STARTRAK were developed to up 457.2 cm, it is difficult to be sure of the accuracy of those equipments, especially in a long range. We also observed that the tracker-based system was affected by another equipment (Wireless LAN Access Point) which uses RF signal even though the influence was not great. In this experiments, the carrier frequency of FASTRAK and a wireless LAN access point were 12.019 KHz and 2.4 GHz, respectively. Even though the vision-based system has a drawback due to an environmental illumination, our system operates well in the normal indoor illumination conditions, irrespective of other environmental conditions.



(a) the real distance (m)



(b) the measurement error (m)



(c) standard deviation (m)

Figure 10. Performance Comparison between Tracker-based system (FASTRAK) and Vision-based System

7. DISCUSSIONS

In this paper, we proposed a simple and robust 3D kendo game which is composed of a multimodal interface (vision-based 3D interface & speech-based interface) and an intelligent agent for perception and reaction between human and cyber-fighter. Reliable, real-time, 3D tracking of humans is a difficult task by using vision-based interface. However, it can allow full body interaction of user with virtual environment to interact with cyber-fighter without distracting the user. Even though users need to be familiar with some limitations to cope with the environment, the suggested system with autonomous fighter can improve real time interactions between human and an opponent fighter in virtual environment and can enhance the life-like impression of fighting actions. From now on, integration of each part gives a really good impression of this system with excitement and very funny fighting with opponent fighter.

A remaining challenge is to develop the kendo system exploiting the force feedback. The force feedback combined with the vision and speech will provide kendo fighting with a full immersion. In addition, a photo-realistic avatar in the 3D virtual environment will provide more realistic experience. In addition, networked VR game with real person will be another challenging task.

REFERENCES

1. Bates J. Aitucher J., Hauptman A., Kantrwoitz M., Loyall A.B., Murakami K., Olbrich P., Popovic Z., Reilly W.S., Sengers P., Welch W., Weyharauch P. and Witkin A., "Edge of Intention", *SIGGRAPH-93 Visual Proceedings*, Machine Culture, *ACM SIGGRAPH*, pp. 113-114, 1993.
2. Fisher S.S., Girard M. and Amkraut S., Menagerie, "Tomorrow's Realities," *SIGGRAPH-93 Visual Proceeding, ACM SIGGRAPH* 1993, pp.212-213, 1993.
3. S-Y Yoon, R. C. Burke, B. M. Blumberg, G. E. Schneider, "Interactive Training for Synthetic Characters", submitted to *AAAI 2000*.
4. Pattie Maes, Trevor Darrell, Burce Blumberg, Alex Pentland, "The ALIVE System: Wireless, Full-body Integration with Autonomous Agents", *In the ACM Special Issue on Multimedia and Multisensory Virtual Worlds*, 1996.
5. Jungwon Yoon, Hyuck-kee Lee, and Jeha Ryu, "Network Gum-Do Simulation Game Using 6 dof Haptic Joysticks", *32nd International Symposium on Robotics (ISR2001)*, Seoul, Korea, April 19-21, 2001, pp. 1777-1782.
6. A. Mulder, Human movement tracking technology, Simon Fraser University: Technical Report 94-1, 1994.
7. W. Woo and Y. Iwadate, "Object-oriented hybrid segmentation using stereo images," *in Proc. SPIE VCIP*, pp. 487-495, Jan. 2000.
8. W. Woo, N. Kim, and Y. Iwadate, "Object segmentation for z-keying using stereo images," *in Proc. WCC*, pp. 1249-1253, Aug. 2000.
9. N. Kim, W. Woo, and M. Tadenuma, "Photo-realistic 3d virtual environment using multiview video," *in Proc. SPIE VCIP*, Jan. 2001.
10. Polhemus, http://www.polhemus.com/ftrakds.htm