3D Direct Manipulation Interface by Human Body Posture and Gaze

Takaki, Kazuya
Department of Intelligent Systems, Kyushu University

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Yonemoto, Satoshi
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of International Conference on Human-Computer Interaction, 2005-07
バージョン：
権利関係：

3D Direct Manipulation Interface by Human Body Posture and Gaze

Kazuya Takaki, Daisaku Arita*, Satoshi Yonemoto** and Rin-ichiro Taniguchi**

* Department of Intelligent Systems, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka 816-8580, Japan
{kazuya_t, arita, rin}@limu.is.kyushu-u.ac.jp

** Department of Intelligent Informatics, Kyushu Sangyo University
2-3-1, Matsukadai, Higashi-ku, Fukuoka 813-8503 Japan
yonemoto@is.kyusan-u.ac.jp

Abstract

Use of 3-D human motion sensing without physical restrictions is the most promising approach to realize seamless coupling between virtual environments and the real world. Motion capturing without any specific markers by computer vision techniques is the most appropriate for such purposes. As the first step, we have developed an avatar motion control by user body postures, and we have applied it to 3D object manipulation in virtual environments. However, the biggest problem here is that the intention of the user can not be fully recognized only by human body postures, and that, sometimes, unintentional human motions cause unintended object manipulations and incorrect selection of the target object.

Currently, we are introducing the user's gaze to control 3D direct manipulation user interface. Using the gaze it becomes possible to control the user interface more accurately and efficiently. In this paper, we outline our gaze detection method based on computer vision techniques and its use to user interface after reviewing our framework of 3D direct manipulation user interface.

1 Introduction

Use of 3-D human motion sensing without physical restrictions is the most promising approach to realize seamless coupling between virtual environments and the real world. Motion capturing without any specific markers by computer vision techniques is the most appropriate for such purposes (Wren, Azarbayejani, Darrell & Pentland, 1997). Many researchers have analyzed 3D human motion from off-line video sequences or real-time image streams (Metaxas, 1997) (Nunomaki, Yonemoto, Arita, Taniguchi & Tsuruta, 2000) (Deutscher, Davison & Reidet, 2001), but an alternative solution to usual motion capture devices is not found yet. Therefore we focus on employing a simplified model of human bodies, which can be stably estimated. In general, such model is defined as color blobs or motion segments. The unwired human motion sensing is applicable for interactive scene such as man-machine interaction. In particular, it is very effective to estimate 3D body postures, or 3D positions of head, arms, feet, etc. because 3D body postures can reflect a user's intention directly.

According to the above consideration, we have developed an avatar motion control by user body postures, and we have applied it to 3D object manipulation in virtual environments (Yonemoto, Arita & Taniguchi, 2000). However, the biggest problem here is that the intention of the user can not be fully recognized only by human body postures, and that, sometimes, unintentional human motions cause unintended object manipulations and incorrect selection of the target object. Here, we introduce the user's gaze to control 3D Direct Manipulation User Interface. Using the gaze it becomes possible to control the user interface more accurately and efficiently.

2 3D Direct Manipulation User Interface

The motion capture based on computer vision technique requires a lot of PCs and cameras, and we can use the system in the limited area where the system is located. In order that we can use the motion capture in the various places, we have developed a new motion capture system which uses a small number of human body features which can be detected stably. It requires only a PC and two IEEE1394-based digital cameras, and works anywhere, under the assumption that the user sit in the front of the system and that it acquires only his/her upper body motion. Then,

we have developed 3D Direct Manipulation User Interface which can make us easy to manipulate the virtual environment, using the motion capture system. Here we introduce the 3D Direct Manipulation User Interface.

2.1 Observing human body postures

From the viewpoint of human posture observation, our method is based on motion synthesis from a limited number of perceptual cues, which can be stably estimated by vision process. Our algorithm consists of the following steps:

- We employ skin color regions of a face and both hands in the image as visual features.
- When a 2D blob is detected in two views, or in multiple views, the 3D position of the blob can be calculated by stereo vision.
- From the 3D positions of user's head and both hands, using model-based analysis, the user's body posture is estimated. The human figure model employed here is shown in Figure 2.

Fig.1 shows an outline of our system and examples of 2D blob tracking results.

In the virtual environments, basic postures of the avatar, which are acquired by observing human body postures, are represented and interpreted as manipulations of the virtual objects (see Figure 3). In addition, we use virtual scene context as a priori knowledge. We assume that virtual objects in the virtual environment can afford avatar's action by simulating the idea of affordance in the real world. In other words, the virtual environments provide action information for the avatar, such as properties of the virtual objects. An important point here is that we can consider scene constraints in the virtual scene to generate more realistic motion beyond the limitation of the real world sensing. Every task in the manipulation is strongly related to objects in the virtual environments.

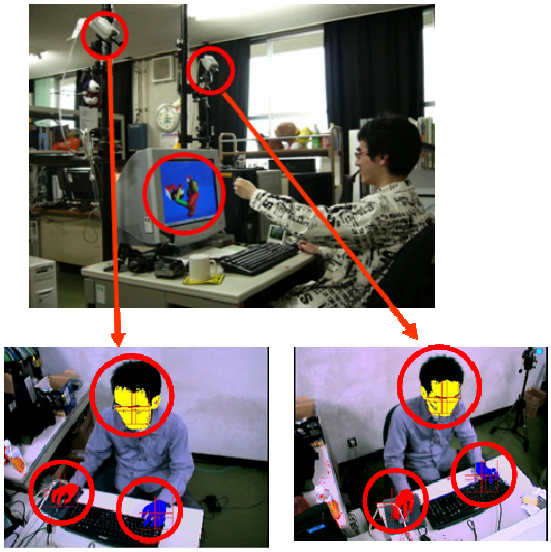

Figure 1: Outline of 3D direct manipulation user interface

2.2 Framework of interaction based on human body posture

We describe our framework to utilize virtual scene contexts as a priori knowledge (Yonemoto & Taniguchi, 2003). In order to make the virtual scene more realistically beyond the limitation of the real world sensing, we can augment the reality in the virtual scene by simulating various events of the real world. We assume that each virtual object affords additional information about user's action by simulating the idea of affordance in the virtual environments. Interaction among the virtual objects and the user (i.e., avatar) should be properly performed by making each virtual object give rise to the related actions. Both the states of the virtual objects and the states of the avatar are picked up whether the virtual objects are handled or not, for example, according to the distance between the virtual objects and the user. For simple example, in grasping a virtual cup, finger motions are afforded as the related actions. Then the

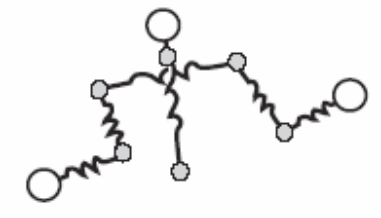

Figure 2: The upper body geometry

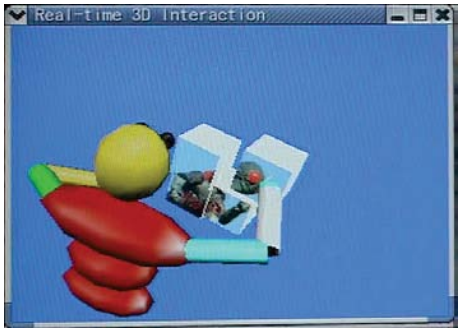

Figure 3: Example of interaction scene

state of the object is changed static into move. These motions are not acquired by the real-world sensing but provided from the virtual object for the purpose of augmenting the virtual scene. Moreover, our system can consider the scene constraints in the virtual environments to simulate the scene events realistically. Since the virtual scene is completely recognized by the system, to reason the scene contexts in the interaction through the virtual environments is not more serious than in the real world sensing.

Table.1 shows the driving methods for the upper body of the avatar. In this table, the upper three rows indicate several parts which should be driven by vision. The lower rows indicate the rest parts driven by motion synthesis. The method **BT** and **GI** means blob tracking and grasp identification respectively. The joint positions with **BT** are described as white circles in Figure 2. **PH** and **AF** mean physically-constrained motion synthesis and afforded motion synthesis respectively. The joint positions with PH are described as gray circles in Figure 2. Detailed motions in object manipulation are driven as the afforded actions.

Table 1: The driving methods of the upper body.

Body part	Method	Category
head position	**BT**	Vision
hand positions	**BT**	Vision
hand states	**GI**	Vision
shoulders	**PH**	motion synthesis
elbows	**PH**	motion synthesis
neck, torso	**PH**	motion synthesis
face direction	**AF**	motion synthesis
fingers	**AF**	motion synthesis

3 Interface with Gaze

The user interface described above uses only three visual features of a human body, it can produce a little and simple manipulation such as grasping, moving etc. To use the gaze information in a user interface, detecting gaze itself is quite important, the gaze information is said to express user's intention generally. Therefore, we discuss using the gaze information for the new input information of the user interface, in order to perform more complicate and advanced manipulations. Here, we introduce how to estimate user's gaze direction and how to use the gaze information in the user interface.

3.1 Gaze detection by vision processing

We develop a gaze measuring system based on vision processing, which does not require any attached sensors and which provides a natural way of gaze sensing. We use a stereo camera only for the gaze detection, and the processing step of gazing sensing is summarized as follows:

- Face direction estimation based on image features (Sako, Whitehouse, Smith & Sutherland, 1994) (Ishii, Arita & Taniguchi, 2003), which is used to transform of face-based local coordinates into the world coordinates for gaze calculation
- Detection of irises, which is realized by extracting circular features around possible eye regions
- Matching the detected irises with our eyeball model, the rotational parameters of the eyeballs are calculated and, then, the gaze is estimated

Here, we describe a detailed method of the gaze detection of our system.

3.1.1 Face direction estimation

We use mouth ends and inner corners of user's eyes for the facial feature points, because they are detected easily and stably. At first we detect ellipses as eye and mouth areas from stereo face images (see Figure 4)(Fitzgibbon & Fisher, 1995). The feature points are located in the eye or mouth ellipse areas, and we detect them using template matching in the ellipse corners. Figure 5 indicates detected face feature points: red ellipses are eye areas, blue dots are the mouth ends, and a yellow dot is the center of the mouth. Then, we calculate the 3D positions of the feature points referring to the results in both views. The face direction is calculated from the 3D positions of the feature points: mouth ends and inner corners of both eyes.

Figure 4: Stereo image pair for gaze detection.

Figure 5: Detected face features.

3.1.2 Gaze direction and Use's viewpoint estimation

Irises are circular shaped, and we detect irises in the eye areas using ellipse detection at first: we get edges of dark area using canny filter, and the center of iris based on ellipse detection (see Figure 6). Then, we estimate the user's

gaze direction and viewpoint. We estimate the 3D position of the eyeball by fitting a simple eyeball model (see Figure 7). In the estimation process, we assume as follows:

- The eyeball is sphere, and its radius is 12mm.
- The center of the eyeball is located on the back of face direction from the center of eye area.
- The center of iris is located on the surface of eyeball.

We can calculate the distance between the center of the eyeball and the center of the eye area easily using a triangle which is made by the center of the eye area, the center of the eyeball and the center of the iris, and we can estimate the center of the eyeball easily. We decide the direction to the center of the iris from the center of the eyeball as the user's gaze direction.

Finally, we estimate the user's viewpoint on the display and also view point in the virtual environment, referring to the gaze direction and the position of the display measured previously. We estimate the virtual viewpoint on the virtual projection plane, and determine the direction to virtual viewpoint from the position of virtual camera as the virtual view direction (see Figure 8).

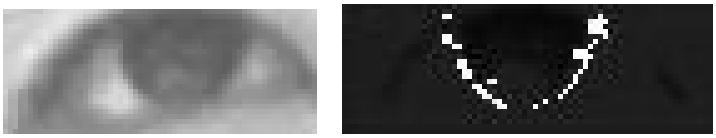

Figure 6: Example of detected iris

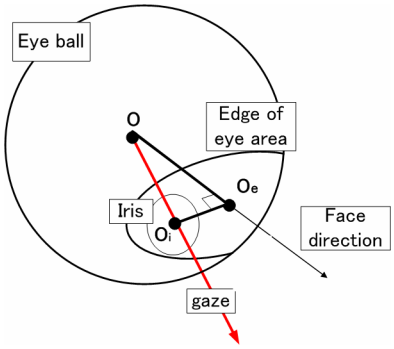

Figure 7: Eyeball model

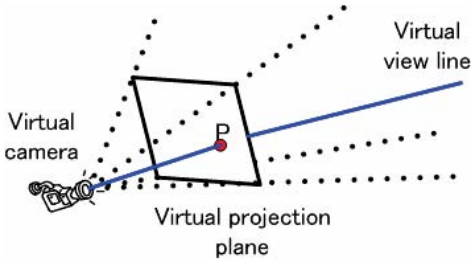

Figure 8: User's view line on virtual environment

3.1.3 Experiments

We have used two cameras and a PC in this experiment. Cameras used here are IEEE1394-based digital cameras, Sony DFW-V500, whose spatial resolution is 640x480 pixels. We have used the 19 inch LCD display. The user was sitting about 50 cm in front of the display. As a primary investigation, we have examined the performance of iris detection, which is the most essential part of gaze detection. Although we have currently achieved detection rate with more than 95%, accuracy of iris shape parameters are not very high. This is due to relatively low spatial resolution of input images. If the accuracy of estimating the gaze direction is higher, the accuracy of estimating the user's intention is higher, so we have to decrease the error.

3.2 Gaze utilization

To make the interface more efficient and more natural, we consider the following three kinds of gaze utilization:

3.2.1 Assistance of manipulation

When we can detect the target object which the user looks at, we can attach annotation to the object, which helps the user's manipulation. The possible annotation is as follows (see Figure 9):

- Enhance the visibility of the target object, such as high lighting, enlargement, or translation (pop-up) of the object. When the user look an object, the system recognizes that the user wants to manipulate the one. Then, the system enhances the visibility of the target object to the grade with which virtual environment does not fail, in order that the user can manipulate it more easily.
- Display supplemental information of the target object, such as properties, possible manipulations, etc. When the system recognizes that the user is confusing by referring to his/her gaze movement, it also displays some assistance information. For example, when the user does not look the objects and the user is confusing (the user's gaze is moving frequently), the system displays the information such as the objects he/she can manipulate. When the user looks at an object (the user's gaze is not moving) and he/she doesn't perform any action, the system displays the information such as how to manipulate the target object.

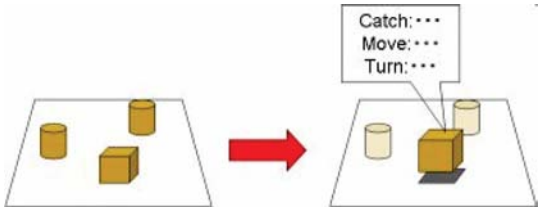

Figure 9: Assistance of manipulation

3.2.2 Disambiguation

In the interface by human posture, the target object is basically decided by the positions of the human hands. However, only by the hand position, it is not easy for the system to distinguish whether the hand is intentionally approaching the object for manipulation or the hand is accidentally approaching. Usually, the target object is located along the line of sight, and the gaze can give information whether the user is going to manipulate the object. Combining the human body motion and the gaze, we can disambiguate possible interpretation of human actions, and as a result, the number of objects and manipulations which can be handled efficiently is increased.

The examples of the disambiguation are as follows:

- The intention of the target object
  When there are multiple of objects on the virtual plane, the user may manipulate unexpected objects. If the user's intention what he/she expects is detected, the system targets only the one (see Figure 10).
- The intention of the manipulation
  When the user wants to manipulate an object, the way to manipulate it changes depending on the purpose of the manipulation. For example, when the user wants to have a cup of tea, he/she grasps the grip of the cup, but when the user wants to see it in detail, he/she may grasp the side of the cup. When the system recognizes the intention of the user, he/she can grasp the cup suitably without other recognition such as hand shape.

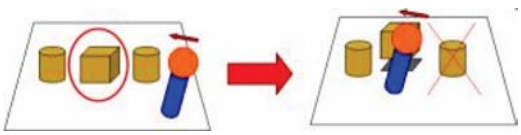

Figure 10: Disambiguation

3.2.3 Concealment of system delay

One of the problems of the vision-based interface is its delay due to vision processing, and the feedback of the human action displayed on a monitor is a bit delayed. This sometimes causes unnatural feeling in using the interface. If we can recognize the user's intention referring to the gaze and if we can predict the user's action, we can compensate the feedback, which can virtually hide the system delay (see Figure 11).

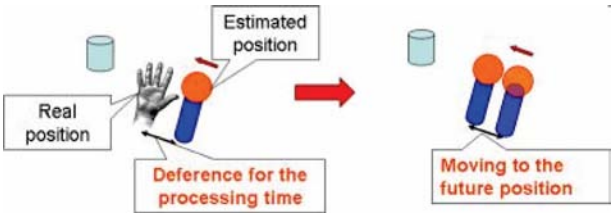

Figure 11: 3.2.3 Concealment of system delay

4 Conclusion

Here, we have introduced the user's gaze to a virtual object manipulation interface in order to control the user interface more accurately and efficiently. The gaze was measured by a computer vision technique, which provides a natural way of sensing. Using the gaze, we can achieve (a) disambiguation of objects and manipulations, (b) assistance of input, (c) concealment of the system delay based on the action prediction. Currently, we are developing a prototypical user interface, and we have to sophisticate it by improving algorithms of vision processing and the user's psychological state recognition.

References

Deutscher, J., Davison, A., & Reidet, I. (2001). Automatic Partitioning of High Dimensional Search Spaces Associated with Articulated Body Motion Capture. *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2, 669-676.

Fitzgibbon, A. W., & Fisher, R. B. (1995). A Buyer's Guide to Conic Fitting. *Proceedings of 5th British Machine Vision Conference* (BMVC), 513-522.

Ishii, S., Arita, D., & Taniguchi, R. (2003). Real-time Head Pose Estimation with Stereo Vision. *Proceedings of the 9th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV2003)*, 79-97.

Metaxas, D. (1997). *Physics-based Deformable Models Applications to Computer Vision, Graphics and Medical Imaging*, Kluwer Academic Publishers.

Nunomaki, T., Yonemoto, S., Arita, D., Taniguchi, R., & Tsuruta, N. (2000). Multi-part Non-rigid Object Tracking Based on Time Model-Space Gradients. *Proceedings of Workshop on Articulated Motion and Deformable Objects (AMDO)*, 72-82.

Sako, H., Whitehouse, M., Smith, A., & Sutherland, A. (1994). Real-Time Facial-Feature Tracking Based on Matching Techniques and Its Applications. Proceedings of International Conference on Pattern Recognition (ICPR) 2, 320-324.

Wren, C., Azarbayejani, A., Darrell, T., & Pentland, A. (1997). Pfinder: Real-Time Tracking of the Human Body, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (7), 780–785.

Yonemoto, S., Arita, D., & Taniguchi, R. (2000). Real-Time Human Motion Analysis and IK-based Human Figure Control. *Proceedings of Workshop on Human Motion*, 149-154.

Yonemoto, S., & R. Taniguchi. (2003). Direct Manipulation Interface with Vision-based Human Figure Control. *Proceedings of HCI International 2003*, 811-815.