A Natural Interface to a Virtual Environment through Computer Vision-estimated Pointing Gestures

Thomas B. Moeslund, Moritz Störring, and Erik Granum

Laboratory of Computer Vision and Media Technology
Aalborg University, Niels Jernes Vej 14
DK-9220 Aalborg East, Denmark
Email:{tbm,mst,eg}@cvmt.auc.dk

Abstract. This paper describes the development of a natural interface to a virtual environment. The interface is through a natural pointing gesture and replaces pointing devices which are normally used to interact with virtual environments. The pointing gesture is estimated in 3D using kinematic knowledge of the arm during pointing and monocular computer vision. The latter is used to extract the 2D position of the user's hand and map it into 3D. Off-line tests show promising results with an average errors of 8*cm* when pointing at a screen 2*m* away.

1 Introduction

A virtual environment is a computer generated world wherein everything imaginable can appear. It has therefore become known as a virtual world or rather a virtual reality (VR). The 'visual entrance' to VR is a screen which acts as a window into the VR. Ideally one may feel immersed in the virtual world. For this to be believable a user is either to wear a head-mounted display or be located in front of a large screen, or even better, be completely surrounded by large screens.

In many VR applications [4] the user needs to interact with the environment, e.g. to pinpoint an object, indicate a direction, or select a menu point. A number of pointing devices and advanced 3D mouses (space mouses) have been developed to support these interactions. These interfaces are based on the computer's terms which many times are not natural or intuitive to use.

In this paper we propose to replace such pointing devices with a computer vision system capable of recognising natural pointing gestures of the hand. We choose to explore how well this may be achieved using just one camera and we will focus on interaction with one of the sides in a VR-CUBE, see figure 1 a), which is sufficient for initial feasibility and usability studies.

2 The Approach

The pointing gesture belongs to the class of gestures known as *deictic gestures* which MacNeill [3] describes as "gestures pointing to something or somebody either concrete or abstract". The use of the gesture depends on the context and the person using it [2]. However, it has mainly two usages: to indicate a direction or to pinpoint a certain object.

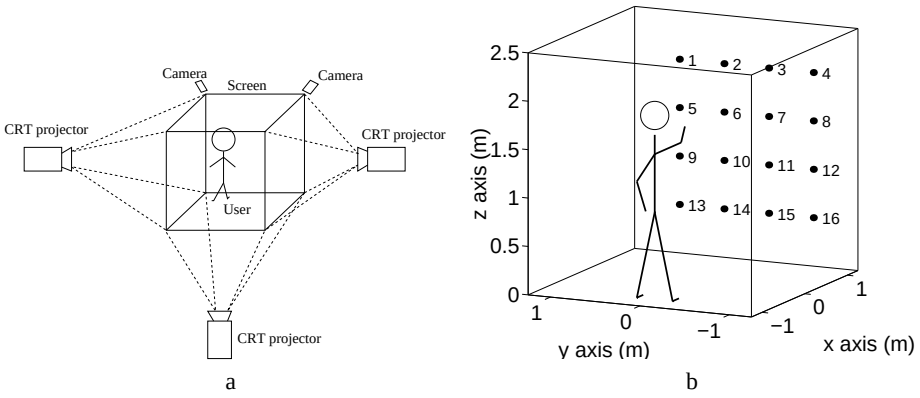

Fig. 1. VR-CUBE: a) Schematic view of the VR-CUBE. The size is 2.5 x 2.5 x 2.5*m*. b) Experimental setup. 16 points in a .5*cm* raster are displayed.

When the object pointing to is more than approximately one meter away, which is usually the case when pointing in a virtual environment, the pointing direction is indicated by the line spanned by the hand (index finger) and the visual focus (defined as the centre-point between the eyes). Experiments have shown that the direction is consistently (for individual users) placed just lateral to the hand-eye line [5].

The user in the VR-CUBE is wearing stereo-glasses and a magnetic tracker is mounted on these glasses. It measures the 3D position and orientation of the user's head which is used to update the images on the screen from the user's point of view. The 3D position of the tracker can be used to estimate the visual focus and therefore only the 3D position of the hand needs to be estimated in order to calculate the pointing direction.

Since we focus on the interaction with only one side we assume that the user's torso is fronto-parallel with respect to the screen. That allows for an estimation of the position of the shoulder based on the position of the head (glasses). The vector between the glasses and the shoulder is called the displacement vector in the following. This is discussed further in section 3.

The 3D position of the hand can usually be found using multiple cameras and triangulation. However, experiments have shown that sometimes the hand is only visible in one camera. Therefore we address the single-camera problem. We exploit the fact that the distance between the shoulder and the hand (denoted R), when pointing, is rather independent of the pointing direction. This implies that the hand, when pointing, will be located on the surface of a sphere with radius R and centre in the user's shoulder.

The camera used in our system is calibrated [6] which enables us to map an image point (pixel) to a 3D line in the VR-CUBE coordinate system. By estimating the position of the hand in the image we obtain an equation of a straight line in 3D and the 3D position of the hand is found as the point where the line intersects the sphere.

2.1 Estimating the 2D Position of the Hand in the Image

The following is done to segment the user's hand and estimate its 2D position in the image. Firstly the image areas where the user's hand could appear when pointing are estimated using the 3D position and orientation of the user's head (from the magnetic tracker), a model of the human motor system and the kinematic constraints related to it, and the camera parameters (calculating the field of view). Furthermore, a first order predictor [1] is used to estimate the position of the hand from the position in the previous image frame.

The histogram of the intensity image can be approximated by bimodal distribution, the brighter pixels originate from the background whereas the darker originate from the user. This is used to segment the user from the background. The optimal threshold between the two distributions can be found by minimising the weighted sum of group variances [4].

The colour variations in the camera image are poor. All colours are close the the gray vector. Therefore the saturation of the image colours is increased by an empirical factor. The red channel of the segmented pixels has maxima in the skin areas as long as the user is not wearing clothes with a high reflectance in the long (red) wavelengths. The histogram of the red channel can be approximated as a bimodal distribution, hence it is also thresholded by minimising the weighted sum of group variances. After thresholding the group of pixels belonging to the hand can be found [4].

3 Pointing Experiments without Visual Feedback

Five users were each asked to point to 16 different points displayed on the screen, see figure 1 b). No visual feedback was given during these experiments, hence the users should be unbiased and show a natural pointing gesture. An image of each pointing gesture was taken together with the data of the magnetic head tracker. The displacement vector between the head tracker and the shoulder was measured for each user.

Figure 2 a) shows the results of a representative pointing experiment. The circles ($\circ$) are the real positions displayed on the screen and the asterisks ($*$) connected by the dashed lines are the respective estimated positions where the user is pointing to. The error in figure 2 a) is up to $0.7m$. There are no estimates for the column to the left because there is no intersection between the sphere described by the hand and the line spanned by the camera and the hand of the user.

The error is increasing the more the user points to the left. This is mainly due to the incorrect assumption that the displacement vector is constant. The direction and magnitude of the displacement vector between the tracker and shoulder is varying.

For each user a lookup table (LUT) of displacement vectors as a function of the head rotation was build using the shoulder position in the image data and the tracker data. Figure 2 b) shows the result of a representative pointing experiment (same as used in figure 2 a) using a LUT of displacement vectors to estimate the 3D position of the shoulder. Notice that after the position of the shoulder has been correction estimates for the left column is available. In figure 2 c) the average result of all experiments are shown. Each inner circle illustrates the average error while each outer circle illustrates the maximum error. The total average is $76mm$ and the maximum error to $30mm$.

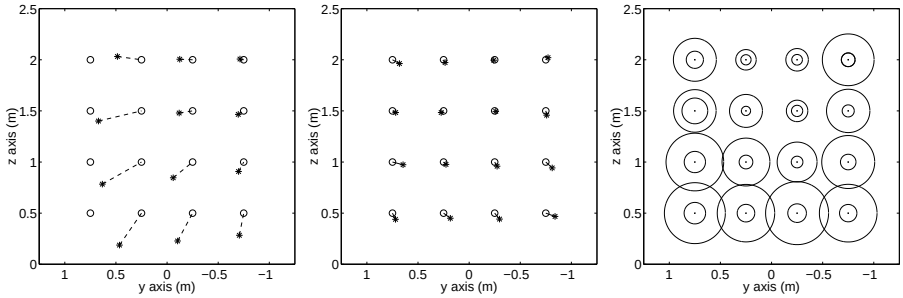

Fig. 2. Results from pointing experiments. See text.

4 Discussion

In this paper we have demonstrated that technical interface devices can be replaced by a natural gesture, namely finger pointing. During off-line tests we showed the average error to be 76*mm* and the maximum error to 308*mm*. This we find to be a rather accurate result given the user is standing two meters away. However, whether this error is too large depends on the application.

In the final system the estimated pointing direction will be indicated by a bright 3D line seen through the stereo glasses starting at the finger of the user and ending at the object pointed to. Thus, the error is less critical since the user is part of the system loop and can correct on the fly.

Currently we are deriving explicit expressions for the error sources presented above. Further experiments will be done in the VR-CUBE to characterise the accuracy and usability as soon as the real time implementation is finished. The experiments will show whether the method allows us to replace the traditional pointing devices as is suggested by our off-line tests.

References

1. Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, INC., 1988.
2. E. Littmann, A. Drees, and H. Ritter. Neural Recognition of Human Pointing Gestures in Real Images. *Neural Processing Letters*, 3:61–71, 1996.
3. D. MacNeill. *Hand and mind: what gestures reveal about thought*. University of Chicago Press, 1992.
4. M. Störring, E. Granum, and T. Moeslund. A Natural Interface to a Virtual Environment through Computer Vision-estimated Pointing Gestures. In *Workshop on Gesture and Sign Language based Human-Computer Interaction*, London, UK., April 2001.
5. J. Taylor and D. McCloskey. Pointing. *Behavioural Brain Research*, 29:1–5, 1988.
6. R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.