VDESIGN: TOWARD IMAGE SEGMENTATION AND COMPOSITION IN CAVE USING FINGER INTERACTIONS

Xiaoming Nan, Ziyang Zhang, Ning Zhang, Fei Guo, Yifeng He, and Ling Guan

Department of Electrical and Computer Engineering, Ryerson University, Canada

ABSTRACT

The Cave Automatic Virtual Environment (CAVE) system is a fully immersive virtual reality system, which can provide users with a realistic experience and a large freedom of interactions. In this paper, we propose *vDesign*, a CAVE-based virtual design environment using finger interactions. Specifically, we focus on the function of image segmentation and composition in the *vDesign* system. In *vDesign*, the user wears a marker on each hand. The interactions of the user are triggered based on the real-time positions of the markers. We design multiple finger interactions for image segmentation and image composition. In image segmentation, the user can use the right finger to select the interested object and the left finger to select the unrelated background. Based on the user's selection, a graph-cut based image segmentation is employed to extract the interested object from the image. In image composition, the user can move, rotate, and scale the segmented objects with fingers and combine them together into a final image. We implemented the *vDesign* prototype and conducted experiments to compare the finger interactions and the traditional wand interactions. The experimental results demonstrated that the proposed finger interactions can provide faster and more accurate interactions compared to the traditional wand interactions.

Index Terms— Cave Automatic Virtual Environment (CAVE), virtual design, image segmentation, image composition, finger interactions.

1. INTRODUCTION

Cave Automatic Virtual Environment (CAVE) is a fully immersive Virtual Reality (VR) system, in which the user can not only perceive the Three-Dimensional (3D) environment but also interact with the virtual objects. The conventional interaction tool in CAVE is the wand, which is a 6 Degrees-Of-Freedom (DOF) interaction tool. However, the wand interaction is neither natural nor intuitive. It requires a learning process, such as memorizing button functions. Aiming for a natural Human Computer Interaction (HCI), we propose finger interactions in the CAVE. With finger interactions, the user in the CAVE can perform a variety of actions, such as menu navigation and object manipulations, in an intuitive way.

We plan to develop a series of CAVE-based applications using finger interactions. We call the whole project using finger interactions *vProject*. Our *vProject* umbrella contains virtual gaming (vGame), virtual presenting (vPresent), virtual working (vWork), virtual designing (vDesign), and others. In vGame, we will develop a CAVE-based squash game, in which the user hits a squash ball with a virtual racket against the walls, which return the ball back to the user. The moving direction of the ball is determined by the touch angle between the ball and the racket, and the moving speed of the ball is determined by the speed of the racket. In vPresent, we will develop a cloud based 3D virtual presenting scheme for interactive product customization, in which the product specialist can show the 3D virtual product in the CAVE and dynamically customize the product based on the feedbacks from the customers, while the customers can provide their opinions in real time when they are viewing a 3D visualization of the product. The proposed vPresent will be a cloud based system, in which the customers are able to access the customized virtual products from anywhere at any time, via desktop computers or mobile devices. in vWork, we will develop a virtual office, in which the user can do basic office work in CAVE, such as sending or receiving emails, and talking to other colleagues via video calls. In this paper, we will present *vDesign*, which is our first project within the *vProject* umbrella.

With the virtual reality technology, virtual design moves the traditional design process to a 3D virtual environment. Designers can freely investigate the product from different angles and dynamically adjust the appearance of the product. Unlike the traditional 2D-based design, virtual design can provide more realistic experience and a larger freedom of interactions. For these reasons, virtual design has been widely used in interior decoration, architecture design, urban planning, manufacture industry, and so on.

In virtual design, designers usually want to combine all the interested elements from different images. For instance, the interior designer can compose multiple interested objects to the personalized wall paper. Traditionally, the tasks of image segmentation and composition are done with the desktop-based software. However, the desktop-based software has the following limitations: 1) The interface in the desktop-based software is not natural, thus requiring a learning process for a beginner. 2) The desktop-based software cannot provide

978-1-4799-1043-4/13/$31.00 ©2013 IEEE

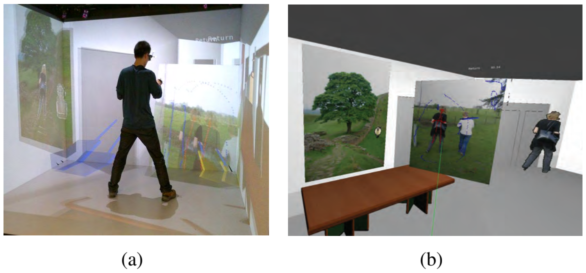

(a) (b)

Fig. 1. *vDesign* in the CAVE: (a) finger interactions performed by the user, and (b) image segmentation and composition in a virtual room.

an immersive environment, without which users cannot get a realistic impression of the design. 3) The 3D manipulations in a desktop computer are not convenient, thus users cannot perceive the precise position of the 3D object.

To enhance the quality of user experience and provide convenient interactions, we propose *vDesign* in this paper. In *vDesign*, the user can perform various tasks using finger interactions, as shown in Fig. 1(a). The user wears a marker on each hand. Each marker is tracked by the tracking system. Based on the real-time tracking data, we design and implement multiple finger interactions for image segmentation and image composition, as shown in Fig. 1(b). In image segmentation, the user uses the right finger to select the interested object, and the left finger to select the unrelated background. Based on the user's selection, a graph-cut based image segmentation [1] is performed to cut the interested object from the image. In image composition, the user can move, rotate, and scale the segmented objects with fingers and compose them together into one image. We implemented *vDesign* prototype and conducted experiments to compare the finger interactions and the traditional wand interactions. The experimental results demonstrated that the proposed finger interactions outperform the conventional wand interactions in terms of time and manipulation distortion.

2. RELATED WORK

In literature, various interactive tools have been used at the CAVE system. Traditionally, the so called Flystick, a wand type remote control, is used with various buttons. Individual button and a combo of buttons correspond to certain actions such as menu selection, directional up/down, etc. Abramyan *et al.* used two types of wands (Nintendo Wii controller and Nunchuk joystick) in angle viewing and manipulation control [2]. Wand was also used in a virtual table tennis game as the hand tracker to mimic the racket in Li *et al.*'s work [3]. Koike and Makino proposed a 3D solid modeling system using wand to draw sketches on the screen, and a 3D model was then converted from the basic sketch [4].

Other researchers proposed various approaches in transferring command from 2D touch screens to the 3D CAVE. Kim *et al.* clicked button and drew arrows on the iPhone/iPod for touch screen to command the CAVE in menu selection and navigation [5]. In data mining, Prachyabrued *et al.* developed an interface based on iPod touch technology, in order to achieve complicated and cluttered 3D data visualization and manipulation in the CAVE environment [6]. Song *et al.* proposed a set of interaction command based on iPod touch, including volume data slicing, drawing, and annotation [7].

Although 2D based interaction utilized recently developed touch screen technology and provided an easy access interface, it still hasn't reached the full capacity that immersive virtual reality provides. Initial works have been proposed in tracking hand and limbs in commanding the CAVE. Kapri *et al.* used marker-based hands and head tracking method for steering-by-pointing in directional control [8]. Flynn *et al.* utilized Microsoft Kinect to track limbs and head in navigating through an office space [9]. Virtual ball juggling using hand motion tracking technique was also reported in literature [10].

3. IMAGE SEGMENTATION AND COMPOSITION IN THE CAVE

3.1. vDesign System Overview

The *vDesign* system is a CAVE-based virtual design environment. In this paper, we focus on the function of image segmentation and composition, which is a component of the *vDesign* system. In *vDesign*, the main menu can be activated by a pull-down action performed by the right finger. Once the main menu appears in front of the user, the user can navigate the menu and choose a function by touching the corresponding menu item, similar to the touch operation in tablets.

If the function of image segmentation and composition is selected, a series of photos will be displayed in the 3D space and the user can pick up one by touching the photo. The selected photo will be automatically enlarged in front of the user for easy operations. The user can draw the strokes on the interested object with the right finger, and on the unrelated background with the left finger, simultaneously. With the strokes as seeds, a graph-cut based image segmentation will be performed to cut the interested object from the image. The user can place additional strokes in an iterative way to improve the segmentation result. To place the segmented object in a proper position on the background image, the user needs to move, rotate, and scale the object with fingers.

3.2. Image Segmentation in the CAVE

Image segmentation is the process of partitioning an image into different regions which may have similar intensity, color or texture [11]. In virtual design, the designer may need a composed image which contains the objects from different sources. To achieve this objective, the first step is to segment the individual objects from different images. In this paper, we employ the graph-cut based image segmentation [12]. Compared with other segmentation methods, the graph-cut based image segmentation can achieve a globally optimal solution. Also, in the graph-cut based image segmentation, the segmented result can be efficiently refined in an iterative way by providing additional seeds on the object and background.

Let $\mathcal{P}$ denote the set of pixels and $A = (A_1, \cdots, A_p, \cdots, A_{|\mathcal{P}|})$ be a binary vector whose component A_p denote the assignment for pixel p ($p \in \mathcal{P}$). Each pixel can be assigned as interested object or unrelated background. User's indications are taken as hard constraints. For undetermined pixels, the cost function [12] can be formulated as follows.

$$E(A) = \lambda \cdot R(A) + B(A), \tag{1}$$

where $R(A)$ is the regional term and $B(A)$ is the boundary term. According to the study in [1], the minimal solution for Equation (1) can be achieved by finding the minimal cut in graph $\mathcal{G}$, whose nodes correspond to pixels in the image. In graph $\mathcal{G}$, the interested object is taken as source node and the background set as sink node. The minimal cut in graph $\mathcal{G}$ is the optimal segmentation with the minimal cost.

We employ the *region-based segmentation accuracy* [13] to evaluate the performance of image segmentation. The *region-based segmentation accuracy* measures the region coincidence between the segmentation result and the ground truth, which is formulated as follows [13].

$$d(S,T) = \frac{|S \bigcap T|}{|S \bigcup T|} = \frac{|S \bigcap T|}{|S| + |T| - |S \bigcap T|}, \tag{2}$$

where $|\cdot|$ is the calculation of the region area, S is the segmented result, and T is the ground truth of interested object. The numerator $|S \bigcap T|$ denotes the coincidence between the segmented result and the ground truth, while the denominator $|S \bigcup T|$ is a normalization factor. The region-based segmentation accuracy $d(S,T)$ is in the range of $[0,1]$. A higher value of $d(S,T)$ indicates a better segmentation.

3.3. Image Composition in the CAVE

Image composition is the process of combining objects from separate images into a final image, to create the illusion that all those objects are parts of the same scene. In *vDesign*, the user first extracts the objects of interest from separate images, and then manipulates the objects to the desired states (e,g., the position, the rotation, and the size) on the final image.

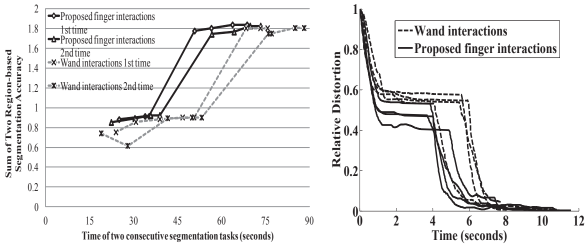
In *vDesign*, object manipulations are performed via fingers. At any time, we can get the 6 DOF tracking data in the format of $(x, y, z, \eta, \theta, \phi)$, for any marker on the hand. The coordinates (x, y, z) represent the position of the marker in the 3D space, and the Euler angles (η, θ, ϕ) represent the rotation of the marker around its local coordinate system. We use the left marker to represent the left index finger and the right marker to represent the right index finger. That is why we call such marker-based interactions *finger interactions*. In *vDesign*, the trigger of an action is determined by the positions of the two markers and the position of the virtual object to be manipulated. For example, a menu item is selected when the distance between the right marker and the center of the menu item is less than a threshold.

In *vDesign*, we define three basic actions for object manipulations, which are moving, rotating, and scaling an object. The moving action is controlled by the midpoint between the two markers. In other words, the object is moved along a path which is parallel to the moving path of the midpoint. The rotating action is determined by four factors: *rotation plane*, *rotation axis*, *rotation direction*, and *rotation angle*. The object manipulation is performed in a discrete-time manner. The *rotation plane* is determined by the two intersected lines: the line (denoted as L_c) passing through the two markers at the current time, and the line (denoted as L_p) passing through the two markers at the previous time. The *rotation axis* is the line perpendicular to the rotation plane and through intersection point. The *rotation direction* is the same to the rotation direction of the two markers. The *rotation angle* is the angle through which line L_p is rotated to coincide with line L_c around the rotation axis along the rotation direction. Scaling in *vDesign* is *uniform scaling*, which means that the object is enlarged or shrunk with the the same *scaling factor* in all directions. The *scaling factor* is defined as the ratio between the distance between the two markers at the current time and that at the previous time.

A composed image consists of multiple objects which may be extracted from other images. The quality of the composed image depends on the final state of each object. We use *manipulation distortion* to measure the manipulation deviations of the objects in the image composition. Let $\mathbf{N}$ denote the set of the objects to be manipulated onto the composed image. Let $\mathbf{V}_n (n \in \mathbf{N})$ denote the set of vertices of object n represented by a 3D polygonal mesh. The *manipulation distortion* $D_n^{(t)}$ of object n at time t, represented by the Mean Squared Error (MSE), is given by $D_n^{(t)} = \frac{1}{|\mathbf{V}_n|} \sum_{i=1}^{|\mathbf{V}_n|} (d_{ni}^{(t)})^2$ where $|\mathbf{V}_n|$ represents the number of the vertices in the set $\mathbf{V}_n$, and $d_{ni}^{(t)}$ represents the distance between the i-th vertex of object n at time t and the corresponding vertex of the object at the target state. The manipulation distortion $D^{(t)}$ of the composed image is the sum of the manipulation distortions of the objects manipulated onto the final image. It is given by $D^{(t)} = \sum_{n \in \mathbf{N}} D_n^{(t)} = \sum_{n \in \mathbf{N}} \frac{1}{|\mathbf{V}_n|} \sum_{i=1}^{|\mathbf{V}_n|} (d_{ni}^{(t)})^2$.

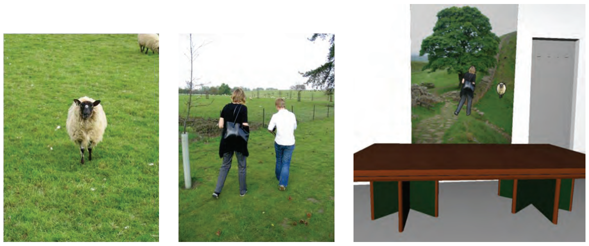
4. EXPERIMENTS

We implemented the prototype of *vDesign* in the CAVE at Ryerson university, Canada. The prototype is developed in C++ language based on the libraries of VR Juggler [14] and OpenSceneGraph [15]. We design image segmentation and



(a) (b)

Fig. 2. Comparison results: (a) comparison of region-based segmentation accuracy in image segmentation, and (b) comparison of manipulation distortion in image composition.



(a) (b) (c)

Fig. 3. Illustration of image segmentation and composition: (a) image 1 with a sheep, (b) image 2 with a lady and a boy, and (c) the composed image with the sheep extracted from image 1 and the lady extracted from image 2.

image composition tasks to evaluate the performance of the proposed *vDesign* system. In the image segmentation task, the user is asked to segment one object from the first image and another object from the second image. The performance of image segmentation is evaluated by the *region-based segmentation accuracy*. The value of region-based segmentation accuracy is calculated online and shown to the user when the user is performing image segmentation in the CAVE. In the image manipulation task, the user needs to move, rotate, and scale the segmented objects and place them onto the specified locations of the background image. *Manipulation distortion* is used to evaluate the performance of image composition. The tested images and ground truth images are selected from the GrabCut image data set of Microsoft Research Cambridge [16]. We conducted user test to compare the performance between the proposed finger interactions and the traditional wand interactions. In the traditional wand interactions, the user can draw strokes on the interested object and background by pressing buttons, and move or rotate the object by moving the wand. We invite an expert user, who has used wand and finger interactions quite frequently, to participate in our test.

We first compare the performance on image segmentation task between the finger interactions and the wand interactions. The user performs the same image segmentation task 2 times with finger interactions and 2 times with wand interactions. We record the time and the region-based segmentation accuracy. The comparison of segmentation accuracy is shown in Fig. 2(a). Observing Fig. 2(a), we can find that the proposed finger interactions perform much better than the wand interactions in terms of convergence time. Compared to the wand interactions, the proposed finger interactions enable quicker and more accurate operations in drawing the strokes on the object and the background. Moreover, the object and the background can be selected simultaneously with finger interactions, while they can only be selected one by one with wand interactions. Therefore, the finger interactions can take less time to achieve the optimal segmentation than the wand interactions. Next, we compare the relative distortion on the image composition task between the proposed finger interactions and the wand interactions. The image composition task includes the manipulations of two objects. The first object extracted from image 1 is manipulated and then placed onto the background image. When the manipulation distortion of the first object is lower than the predefined distortion threshold of 0.008, the user starts to manipulate the second object which is extracted from image 2. The image composition task is repeated 4 times. Fig. 2(b) shows the normalized manipulation distortion of each experiment. From Fig. 2(b), we can see that the proposed finger interactions can provide faster interactions in object manipulations compared to the conventional wand interactions. Image composition is illustrated in Fig. 3. The first object is the sheep, which is extracted from image 1 shown in Fig. 3(a). The second object is the lady, which is extracted from image 2 shown in Fig. 3(b). The two objects are manipulated and then placed onto another background image. The composed image is used as a picture frame hung on the wall of the virtual room, as shown in Fig. 3(c).

5. CONCLUSION

In this paper, we propose *vDesign*, a CAVE-based virtual design system using finger interactions. Specifically, we investigate the function of image segmentation and composition in the *vDesign* system. We implement multiple interactions for image segmentation and image composition. In image segmentation, the user can use the right finger to select the interested object and the left finger to select the unrelated background. Based on the user's selection, a graph-cut based image segmentation is performed to extract the interested object from the image. In image composition, the user can move, rotate, and scale the segmented objects with fingers and then combine them together into a final image. We implemented *vDesign* prototype, via which we conducted experiments to compare the finger interactions and the traditional wand interactions. The experimental results demonstrated that the proposed finger interactions can provide faster and more accurate interactions compared to the traditional wand interactions.

6. REFERENCES

[1] Yuri Boykov and Gareth Funka-Lea, "Graph cuts and efficient N-D image segmentation," *International Journal of Computer Vision*, vol. 70, pp. 109–131, 2006.

[2] Lucy Abramyan, Mark Powell, and Jeffrey Norris, "Stage: Controlling space robots from a cave on earth," in *Proc. of IEEE Aerospace Conference*, 2012.

[3] Yingzhu Li, L-K Shark, Sarah Jane Hobbs, and James Ingham, "Real-time immersive table tennis game for two players with motion tracking," in *Information Visualisation (IV), 2010 14th International Conference*. IEEE, 2010, pp. 500–505.

[4] Mikiko Koike and Mitsunori Makino, "Crayon a 3d solid modeling system on the cave," in *Proc. of IEEE International Conference on Image and Graphics*, 2009, pp. 634–639.

[5] Ji-Sun Kim, Denis Gračanin, Krešimir Matković, and Francis Quek, "The effects of finger-walking in place (fwip) for spatial knowledge acquisition in virtual environments," in *Springer Smart Graphics*, 2010, pp. 56–67.

[6] Mores Prachyabrued, David Ducrest, and Christoph Borst, "Handymap: a selection interface for cluttered vr environments using a tracked hand-held touch device," *Advances in Visual Computing*, pp. 45–54, 2011.

[7] Peng Song, Wooi Boon Goh, Chi-Wing Fu, Qiang Meng, and Pheng-Ann Heng, "Wysiwyf: exploring and annotating volume data with a tangible handheld device," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1333–1342.

[8] Anette von Kapri, Tobias Rick, and Steven Feiner, "Comparing steering-based travel techniques for search tasks in a cave," in *Proc. of IEEE Virtual Reality Conference*, 2011, pp. 91–94.

[9] Carl Flynn, "Visualising and interacting with a cave using real-world sensor data," 2011.

[10] Minato Mizutori, Koichi Hirota, and Yasushi Ikei, "Skillful manipulation of virtual objects: Implementation of juggling in a virtual environment," in *Proc. of IEEE Virtual Systems and Multimedia*, 2012, pp. 79–86.

[11] Shital Raut, M Raghuvanshi, R Dharaskar, and Adarsh Raut, "Image segmentation – a state-of-art survey for prediction," in *Proc. of IEEE International Conference on Advanced Computer Control*, 2009, pp. 420–424.

[12] Yuri Y Boykov and M-P Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. of IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 105–112.

[13] Feng Ge, Song Wang, and Tiecheng Liu, "New benchmark for image segmentation evaluation," *Journal of Electronic Imaging*, vol. 16, no. 3, 2007.

[14] VR juggler, "http://vrjuggler.org/".

[15] Openscenegraph, "http://www.openscenegraph.org".

[16] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, 2004, vol. 23, pp. 309–314.