DepthTouch:
Using Depth-Sensing Camera to Enable Freehand Interactions On and Above the Interactive Surface

Hrvoje Benko and Andrew D. Wilson

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{benko, awilson}@microsoft.com

March 2009
Technical Report
MSR-TR-2009-23

This work has also been presented as a poster presentation at the *IEEE Workshop on Tabletops and Interactive Surfaces '08.* Amsterdam, the Netherlands, October 1-3, 2008.

Abstract

DepthTouch is an exploratory interactive surface system that combines the benefits of multi-touch-sensitive surface with the ability to interact in the hover space in front of the surface. DepthTouch uses a depth-sensing camera, which reports a range value per pixel in addition to color, to track the 3D position of the user's head and hand through a transparent vertical display screen (DNP HoloScreen). The camera location behind the screen minimizes situations in which one hand occludes the other and allows for tracking of the user's interactions on and off the surface by segmentation of range data. We outline our system hardware, the tracking implementation, and 3D interactions enabled. We also discuss feedback from hundreds of preliminary users and implications for future research.

1. Introduction

Most of the interactive touch-sensitive surface systems restrict the user interaction to a 2D plane of the surface and actively disregard the interactions that happen above it. Even the interactive surfaces that support interactions with tangible objects, commonly track such objects only when in contact with the 2D plane, leaving the 3D interaction space above the surface largely underutilized. 3D interactions in the hover space above the surface have previously been explored within augmented and virtual reality fields with the use of tracked gloves, pens or styli, (e.g., [1]). However, much of the appeal of touch-sensitive interactive surfaces is due to the directness of such interfaces, which do not require the user to hold or wear any additional input devices to interact with the displayed content.

Only a handful of interactive surface projects have explored freehand 3D interactions without any physical trackers or markers. Illuminating Clay [3] used a laser-range-sensing technology to facilitate manipulations of a morphpable projected surface. Recently, Wilson demonstrated the use of a depth-sensing camera to support interactive modification of the terrain in a car-driving simulation called *Micromotocross* [6]. He speculated that depth-sensing cameras will enable easier recognition of 3D gestures above the surface. In this paper, we describe a novel system, called *DepthTouch*, which explores such freehand 3D interactions while preserving the "walk-up-and-use" simplicity of a multi-touch interactive surface.

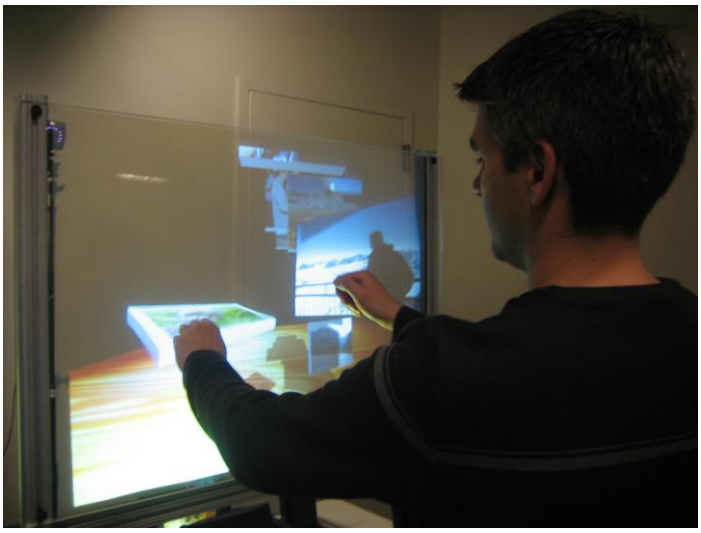

Figure 1. Interacting with DepthTouch: user's left hand is touching an object of interest, while his right hand is adjusting the orientation and depth of that object by moving in mid-air above the surface.

2. DepthTouch

DepthTouch is an interactive prototype that enables a single user to view and interact with a virtual 3D scene projected on a transparent vertical surface (Figure 1). We were motivated by the idea to design a system in which the user can visualize and manipulate 3D virtual objects using freehand gestures on the

interactive surface and above it (in mid-air), while not needing to wear any special tracking or display equipment.

Similar to numerous "fish tank" virtual reality (VR) systems, the DepthTouch prototype tracks the user's head position and renders the correct perspective view of a 3D virtual scene for a single user. However, while such fish tank VR systems typically require the user to wear physical trackers, such as magnetic trackers or visual fiducial markers, DepthTouch relies on a depth-sensing camera to track the user's head and hands, thus allowing for completely tracker-free implementation.

Primarily, we wanted to preserve the "walk-up-and-use" simplicity of the touch-sensitive interactive surface, which does not require the user to acquire any additional tools in order to interact with it. Additionally, we wanted to enable 3D freehand interactions in the hover space above the interactive surface to extend the current touch-based 2D interaction vocabulary.

2.1. System Hardware

DepthTouch consists of a depth-sensing camera (ZSense depth-sensing camera from 3DV Systems, Ltd. [1]), a transparent vertical display screen (DNP HoloScreen) and a "short-throw" projector (NEC WT610, 1024x768 pixel resolution) (Figure 2). In addition to these components, a desktop PC computer is used for processing the camera data and driving the display.

ZSense camera computes a depth-map image (320x240 pixels at 30Hz) by timing the pulsed infra-red light released by the camera and reflected of the objects in front of it: the more light gets returned, the closer the object is at that particular pixel. While ZSense device also contains a separate color camera and is capable of reporting a full "RGBZ" image, in this project, we have not used the color image. For details on how the ZSense camera works, please refer to the descriptions in [1] and [6].

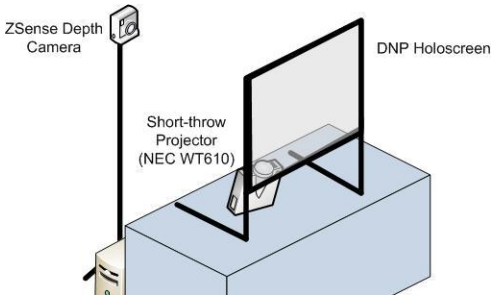

Figure 2. DepthTouch system components.

The motivation behind the use of the transparent screen is both practical and fun: it allows for the depth-sensing camera to be placed directly behind the screen and it further enhances the three-dimensionality of the interface as the surface is not just a 2D plane, but rather a window that looks at a 3D virtual scene embedded in a real world. The camera location behind the screen minimizes situations in which one hand occludes the other and allows for tracking of the user's hands by relatively segmentation of the range data (see Section 2.2).

DepthTouch system shares many configuration similarities with Wilson's TouchLight prototype [7]. TouchLight used two infra-red cameras in the back of

the same transparent screen we used in DepthTouch. While the main idea behind the DepthTouch system is to capture and track 3D interactions on and in-front-of the surface, TouchLight system used the two camera views to capture and track the actions only on the surface, and discarded any interactions at any other depth. The camera setup in TouchLight could have been used to obtain depth information through correlation-based stereo approach; however, this was not desired or implemented.

In comparison with the pulsed light approach used by the ZSense camera, correlation-based stereo has a number of shortcomings: it is hard to match the images in regions where there are no textures, cameras require fine calibration, and the system requires significant amount of processing power to run at interactive rates.

2.2. Tracking by Segmenting the Depth Image

We use the depth image exclusively to track all user interactions on and in front of the surface. The tracking process is relatively simple and consists of four steps.

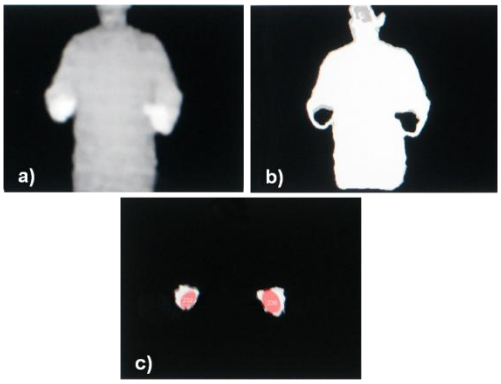

Figure 3. Segmenting the user's body using depth values: a) the depth image of the user standing in front of the screen, b) the torso image shows segmented body without the outstretched hands, and c) the hands image showing segmented arms and hands that can now be easily tracked.

First, we discard the depth data that is either too close or two far from the camera thus ensuring a working depth of about 4 feet in front of the screen. The cleared image is then mapped to the screen coordinates using a projective transform and a simple calibration procedure that maps four corners of the screen to the camera image (similar to [7]). The resulting image can be seen in Figure 3a. While this 2D calibration does not give us a completely orthographic depth image, the perspective differences are minimal given our camera location and the relatively short depth range. This calibration is sufficient for the interactions we explored in this paper, but improvements might be required for more precise interactions.

Second, since most of our depth pixels in the resulting depth image come from the user's torso, we compute the average depth value of the entire image and use that value to segment the user's torso image (Figure 3b) from the hands image which contains the arms and hands in-front of the torso (Figure 3c).

Third, the head position is computed by scanning the torso image and averaging position values of all lit pixels across the top 30 scan lines containing lit

pixels. We check that pixels belong to one contiguous region and that the detected head is at least 30 pixels wide.

Fourth, we use the segmented hands image (Figure 3c) to track the hand blobs and compute their position. We can detect if a hand is touching the surface by comparing its depth to the screen depth at that pixel. The default screen depth at every pixel is obtained during a simple calibration routine by taking a depth image of a large piece of cardboard covering the entire screen.

This four step tracking procedure is simple and robust, but suffers from several limitations. The depth image tends to be noisy, which requires smoothing of depth values and filtering of position estimates using a Kalman filter. This introduces a slight, but noticeable lag. The noise also makes is difficult to reliably track very small extruded points (e.g., fingers) which is why our users interacted mostly with open hands or fists.

Also, we assume that the user's head is the highest part of the detected torso; however, this assumption fails if the user raises their arm above their head or if there are multiple users. While the head-hand confusion errors can be minimized by careful placement of the camera and the screen, a more complex algorithm is needed in order to enable tracking of multiple heads. Since DepthTouch is a single user perspective visualization system, we found this simple algorithm to work remarkably well. Finally, our algorithm is unable to distinguish between hands that are close together, close to the user's body, or not in front of the body (e.g., to the side); however given our setup, such situations are rare.

2.3. Interacting with the Virtual 3D Scene

We used the XNA Game Studio 2.0 platform for our 3D scene visualization. The scene consists of a wood textured horizontal surface above which are several 3D objects. Currently, we use 3D picture frames in our demonstrations (Figure 1), but loading any other 3D models is possible.

The current DepthTouch prototype enables the following three types of interactions: perspective view manipulation based on the user's head position, touch-based 2D interactions in the surface plane, and mid-air freehand 3D interactions above the surface.

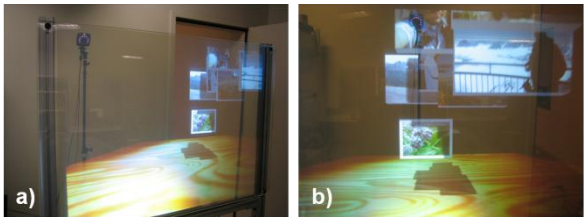

Figure 4. Perspective view of our virtual scene is based on the user's head position: a) a view from the left, b) a view from the middle of the screen.

While we do not provide a stereoscopic view as that would require our user to wear some kind of glasses, we provide a correct perspective 3D view to the user based on the position of their head (Figure 4). In addition to the motion parallax obtained by continuous tracking of the head, we enhance the user's depth perception by providing real-time virtual shadows between the objects and the virtual plane at the bottom of the screen.

The screen also behaves in a manner similar to other multi-touch screens. When the user is touching the object on the screen, they can select it and move it in the surface plane by dragging it around (Figure 5a).

Lastly, we also allow for fine manipulation of the object rotation and depth by performing mid-air interactions with the second hand, while keeping the object selected with the first hand. The object can be rotated in place by moving the second hand in plane above the surface (Figure 5b) or brought closer or further in depth by moving the second hand closer or further away from the user's body (Figure 5c). We do not use the 3D orientation of tracked hand points, and therefore map the object rotation to the simple hand movement in plane. Computing hand orientation with the ZSense camera, while possible, produces rather noisy results due to the noise in the depth values and relatively small detected hand area over which the values can be aggregated.

To select an object and modify it using mid-air interactions, we require the user to explicitly touch and hold the object on the surface with one hand. This "grounding" gesture, while somewhat restrictive, as it requires bimanual operation, eliminates the "Midas touch" problem of mid-air selections.

The space of possible 3D interactions above the surface is far larger than we explored here and these interactions should be considered mainly as a proof-of concept. Our main goal was to explore the feasibility of using depth-sensing cameras for interactive surface applications.

3. Initial User Observations

DepthTouch has been exhibited at a high-traffic location in our organization and hundreds of people of varying ages and backgrounds have interacted with it. Many have commented on the magic-like qualities of being able to interact with the system without wearing any tracking devices. People easily discovered that the system is tracking their head and most were very impressed with the 3D depth perception offered by the motion-parallax and shadowing cues. In fact, many commented that this is really compelling because of the transparent nature of the screen which really reinforced the idea of looking through a window.

We noticed that it was fairly difficult to discover how to properly interact with our system without a demonstration. Partially that was due to the fact that DepthTouch is a single user system and having multiple people in front of it often resulted in resource fighting. Furthermore, it was not obvious how to perform our mid-air gestures. Given the familiarity of our users with the multi-touch technology, most had no problems moving objects around on the surface. However, few expected to be able to do gestures in mid-air, which resulted in frequent unintentional and frustrating modifications of objects' depth or orientation as users moved their other hand while they were engaged with an object. Our design choice of "grounding" 3D gestures required that the user carefully

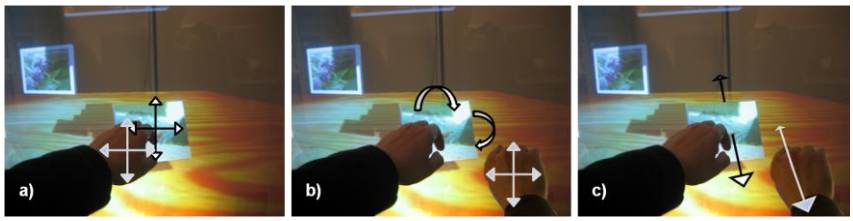

Figure 5. Freehand interactions on and above the surface: a) by touching a surface, the user can select an object of interest and drag it around in plane, b) while keeping the object selected with the first hand, the object can be rotated in place by moving the second hand in plane above the surface, or c) brought closer or further in depth by moving the second hand closer or further away from the user's body.

controls the movement of one hand on the surface and the other one above the surface which is not inherently difficult, but was very unfamiliar to our users. What made it challenging, was the precise movement control of the hand in mid-air, as hands without firm support have a tendency to drift [4].

4. Conclusions and Future Work

This initial user feedback, even while slightly negative, is important as we continue our exploration of above the surface interactions. We believe that while 3D mid-air interactions have a large potential, it is critical to find better ways to delimit the beginning and the end of such gestures. In many ways, this mode selection problem is the critical problem that needs to be solved for 3D gestures to become viable input for interactive surfaces. We offered a solution of "grounding" the user's other hand to signal a mode switch for depth interaction. However, judging from the user feedback, this two-handed approach needs simplification. Other possible selection solutions include a hand pinching gesture [5] or pre-defined hand movements [4] (such as air-tap or crossing into a particular area).

Our initial exploration has demonstrated the feasibility of using a depth-sensing camera to enable a richer set of 3D interactions above the interactive surface in addition to standard multi-touch interactions. In the future, we hope to continue investigating more physics-based interactions to facilitate grasping and placement of virtual objects above and on the surface.

5. Acknowledgement

Thanks to 3DV Systems, Ltd. for providing us the ZSense camera.

References

[1] L.D. Cutler, B. Fröhlich, and P. Hanrahan, "Two-handed Direct Manipulation on the Responsive Workbench", *I3D 1997*, 107-114.

[2] G. J. Iddan and G. Yahav, "3D Imaging in the Studio," *SPIE*, vol. 4298, 2001. 48.

[3] B. Piper, C. Ratti, and H. Ishii, "Illuminating Clay: A 3-D Tangible Interface for Landscape Analysis", *CHI 2002*.

[4] S. Subramanian, D. Aliakseyeu, and A. Lucero, "Multi-Layer Interaction for Digital Tables", *UIST 2006*, 269-272.

[5] A. Wilson, "Robust Computer Vision-Based Detection of Pinching for One and Two-Handed Gesture Input", *UIST 2006*.

[6] A. Wilson, "Depth-Sensing Video Cameras for 3D Tangible Tabletop Interaction", *IEEE Tabletop 2007*, 201-204.

[7] A. Wilson, "TouchLight: An Imaging Touch Screen and Display for Gesture-Based Interaction", *ICMI 2004*, 69-76.