Abstract
Conventional 2D touch-based interaction methods for handheld Augmented Reality (AR) cannot provide intuitive 3D interaction due to a lack of natural gesture input with real-time depth information. The goal of this research is to develop a natural interaction technique for manipulating virtual objects in 3D space on handheld AR devices. We present a novel method that is based on identifying the positions and movements of the user's fingertips, and mapping these gestures onto corresponding manipulations of the virtual objects in the AR scene. We conducted a user study to evaluate this method by comparing it with a common touch-based interface under different AR scenarios. The results indicate that although our method takes longer time, it is more natural and enjoyable to use.

Author Keywords
3D interaction technique; natural gesture interaction; fingertip detection; handheld augmented reality

ACM Classification Keywords

Huidong Bai
The HIT Lab NZ,
University of Canterbury,
Christchurch, 8041 New Zealand
huidong.bai@pg.canterbury.ac.nz

Mark Billinghurst
The HIT Lab NZ,
University of Canterbury,
Christchurch, 8041 New Zealand
mark.billinghurst@hitlabnz.org

Lei Gao
The HIT Lab NZ,
University of Canterbury,
Christchurch, 8041 New Zealand
lei.gao@pg.canterbury.ac.nz

Jihad El-Sana
The Department of Computer Science, Ben-Gurion University of the Negev, Israel
el-sana@cs.bgu.ac.il

Copyright is held by the author/owner(s).
ISMAR'13, October 1 – 4, 2013, Adelaide, Australia.
**Introduction**

In recent years, mobile Augmented Reality (AR) has become very popular. Using a video see-through AR interface on smartphones or tablets, a user can see virtual graphics superimposed on live video of the real world. However, in order for handheld AR to reach its full potential, users should be able to interact with virtual objects to performing translation, rotation or scaling in the AR scene. However, current interaction methods are mainly limited to 2D touch screen pointing and clicking. These suffer from several problems such as finger occlusion of on-screen content, having the interaction area being limited to the screen size, and using 2D input for 3D interaction. Our main motivation is to investigate the potential of 3D gesture interaction as an alternative input technique for handheld AR that overcomes some of these limitations.

**Related Work**

Different types of handheld AR interfaces have been developed over the years, each with their own limitations and best practices. The touch screen is commonly available on the current generation of handheld mobile devices, and touching is the most popular interaction way for current handheld AR applications. However, instead of touching the screen of mobile devices, natural gesture interaction could be an alternative input method. For example, the position and orientation of a finger or palm in midair could be captured by the user-facing camera, analyzed by computer vision algorithms, and the results could be mapped to the virtual scene for manipulation [3] [5]. However, there are some limitations to these previous interaction approaches, such as the lack of accurate depth sensing of fingertips.

In contrast, our interaction technique provides 6 degree-of-freedom (DOF) manipulations using natural finger-based gestures for translating, rotating, or scaling a virtual object in a handheld AR system. Due to current hardware limitations, we chose a tablet to build our prototype, and attached a short range RGB-Depth camera to it. With this implementation, users can easily use their bare fingers to conduct straightforward 3D manipulations of virtual objects in handheld AR applications.

**3D Natural Gesture Interaction**

Using midair gesture interaction methods for handheld AR, the user normally holds the mobile device with one hand while the other free hand can be captured by an RGB-Depth camera and analyzed by computer vision algorithms to find 3D gestures that can control the AR scene. In this section, we design a complete 3D interface for handheld AR applications, in which the user can move, rotate or pinch fingers in 3D space in front of the camera for natural 3D manipulations. In the following paragraphs, we introduce the markerless 3D gesture-based interaction design in more detail.

**Object Selection**

We use a pinch-like gesture with the thumb and index finger to select a virtual target, which is very similar to how people grasp real objects using two fingers. When the midpoint of the two detected fingertips is completely inside the geometric space of the virtual object, the virtual object is selected and becomes one candidate for further possible manipulations. A selection is cancelled by using a common countdown timer method: compared with the previous state, keeping the midpoint of two fingertips relatively still inside a tiny space region (10×10×10mm in our case).
around the current position longer than a certain time (2 seconds in our case), and the object selection state will be canceled.

Canonical Manipulation
The RGB-Depth camera is used so that we can retrieve the depth image synchronously instead of only the RGB frame generated from a normal camera. With this we are able to extract 3D position coordinates of detected fingertips and project them into the AR marker’s world coordinate system, which is also used by the virtual object. As a result, we can directly use the absolute position of two fingertips’ midpoint in space as the 3D geometric center of a virtual object to complete the translation. We also define the rotation input as the absolute 3D pose of the line connecting the user’s thumb and index fingertips in space. Finally we apply a two finger pinch-like gesture and use the distance between the two fingertips as the scaling input.

Finger-based Gesture Detection
We segment the hand region from the RGB frame based on a generalized statistical skin colour model described by Lee and Höllerer [4], and then use the threshold (a distance range between 35cm and 90cm) in the depth image to remove the noise, such as the table in the background with a similar colour to human skin. We use a distance transformation to find a single connected component in the filtrated hand image. Then we use Rajesh’s method [2] to separate the fingers from the whole hand region detected. The point that has the maximum distance value in distance transformation is defined as the center point of the palm. Specifically, the maximum distance D is considered as half of the palm width and twice of one finger width. Then the fingers can be completely eroded by using the element with the width D/2 and to leave only the palm region. For each finger area, we find the minimum rectangle (a rotated rectangle fitted to the finger area) and calculate the midpoints of the four edges of the rectangle. The fingertip position is defined as the midpoint that is farthest from the center point of the hand region (Figure 1).

System Prototype
The prototype runs on a Samsung XE500T1C tablet with a Windows 7 operating system, featuring an Intel(R) Core™ i5 CPU, 4 GB RAM and an Intel(R) HD integrated graphics card. A depth sensor DS325 from SoftKinetic has been used as an external RGB-Depth camera. This provides a color and depth frame both with 320 by 240 resolution to the tablet via a USB connection. However, the sensor requires an indoor operating environment, which means that the output could be influenced by different lighting condition. The sensor also has an operational range of between 15 cm and 100 cm, and it is easy for users to locate their hands from the tablet with such a distance. Thus, we attached the sensor directly on the back of the tablet for our prototyping (Figure 2).

We combined the OpenCV\(^1\) with the OpenNI\(^2\) to obtain image data from the depth sensor, and rendered a virtual scene in our system with the OpenGL\(^3\). AR tracking is implemented using a natural feature-tracking library called OPIRA [1]. We choose this library due to its robustness and fast computation time.

\(^{1}\) http://opencv.org/
\(^{2}\) http://www.openni.org/
\(^{3}\) http://www.opengl.org/
Furthermore, OPIRA has a convenient interface for integrating OpenGL which is easy for scene-rendering.

**User Studies**
To investigate the performance and usability of our manipulation technique, we conducted a user study comparing it with a traditional touch approach across three fundamental scenarios with varying tasks.

**Experimental setup and procedure**
We set up the user study using a within-group factorial design where the independent variables were the manipulation technique and task scenario. The manipulation techniques are our proposed novel natural 3D gesture interaction method and traditional 2D screen-touch input, while the test scenarios contain three different experimental tasks with varying subtasks. The dependent variable is task completion time and we also measured user preferences for both techniques in terms of usability.

To begin the study, each participant was asked to complete a pre-test questionnaire about age, gender and prior experience in touch-based mobile devices, 3D gaming interfaces and mixed or augmented reality. A brief introduction to handheld AR was then given to the participant, followed by a detailed instruction of the 2D touch and 3D gesture manipulations used in our testing environment. The participant learned the general operation attention, basic interface usage, and overall task content. Afterwards, each participant had ten minutes to practice both interaction techniques. Once they started the study, they were not interrupted or given any help. Upon the completion of the practical task with each interaction method, they were asked to fill out a per-condition questionnaire (Table 1) and gave further comments on a post-experiment questionnaire (Table 2) at the end of the evaluation. The whole user study took approximately 45 minutes on average.

For the evaluation, we collected the user preference with seven questions related to the usability in Table 1, on a nine point Likert-scale (1 to 9 with 1 indicating strongly disagree while 9 indicating strongly agree) for each subjective questionnaire item. Furthermore, we configured our testing system to automatically measure the task completion time of the participants.

**Subjects**
32 participants (16 male and 16 female) were recruited from outside of the university for the experiment. Their ages ranged from 17 to 52 years old (M = 35.03, SD = 10.28). All participants were right-handed. During the experimental tests, 29 participants held the device in their left hand and used the right hand for interaction, while for the other three it was the other way around. No significant differences could be observed regarding handedness. All of them used the index finger for touch input and extra thumb finger for gesture manipulations. Although 27 of them used touch screen devices frequently, only six of them had some experience of using 3D interfaces, mainly from the game consoles like Microsoft Kinect and Nintendo Wii. None of them had previous experience with using mixed or augmented reality interfaces.

**Tested Scenarios**
We used the basic canonical manipulation tasks “translation, rotation, and scaling” in the task design, and built three specific scenarios with several subtasks to cover typical 3D manipulation situations in handheld AR applications. To manually identify the desired object
for subsequent manipulation, another canonical operation selection is used.

Each participant was asked to perform several experimental tasks using the two interaction interfaces (traditional 2D touch and novel 3D gesture interaction) respectively. The interfaces were presented in a random order to the participants to exclude potential learning effects. One additional selection and three types of essential manipulations (translation, rotation, and scaling) were included in tasks for each interface test. The order of tasks and related sub-tests were randomized for each participant to avoid any order-related influences on the results.

The experimental task was to select and manipulate a virtual cuboid in a handheld AR application and match it to the indicated target position, pose or size. For all tasks, our system set the target in blue, green for the object the participant hope to control, and red for the selected object which is currently manipulated. All test tasks were presented in the same virtual AR background environment. A black and white textured plane printed on a 297×420mm piece of paper was the target marker for the AR tracking. Meanwhile, both the indicated target and manipulated object were clearly displayed on the same screen, and the participant could inspect the scenario from a different perspective by moving the tablet freely to understand the task before officially conducting the actual test (Figure 3).

Subjects were told to perform the task as fast and accurately as possible. The timing of the task was automatically started after the virtual object was successfully selected, and was stopped automatically when the task was completed, which means that the cuboid was changed to the required status in the space by the participant’s operations.

**Experimental result**
To analyze the performance time and the user questionnaire, we performed a Wilcoxon Signed-Rank Test using the Z statistic with a significance level of 0.01. The Dependent T-Test for analyzing time performance was not applicable because the abnormal distribution of data sets that we collected is confirmed.

Analyzing the data from the performance measurements, we found a significant difference between two interaction methods in terms of overall completion time ($z[n:32] = -4.862, p <0.0001$). On average the tasks were performed significantly faster with 2D screen-touch interaction. When inspecting the mean completion time for each manipulation task, significant differences could also be found for translation ($z[n:32] = -4.020, p <0.0001$), rotation ($z[n:32] = -4.937, p <0.0001$) and scaling ($z[n:32] = -4.619, p <0.0001$). Subjects took more time (around 50 sec.) to finish all tasks with the 3D gesture-based interaction compared to 2D touch-based interaction.

![Figure 3. Experiment setup.](image)

![Figure 4. Users’ average rating for two interfaces.](image)
Analyzing the result of the subjective questionnaire terms from Q1 to Q7 presented in Table 1 for the four manipulations, we got detailed results about users’ preference (Figure 4) and the significant difference (Table 3) between two interaction methods. The results reveal that users thought the 3D gesture was more enjoyable to use, and no significant difference is found between using 3D gesture or 2D touch input in terms of naturalness and mental stress.

<table>
<thead>
<tr>
<th></th>
<th>Translation</th>
<th>Rotation</th>
<th>Scaling</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Z</td>
<td>p</td>
<td>Z</td>
<td>p</td>
</tr>
<tr>
<td>Q1</td>
<td>-3.532</td>
<td>0.000412</td>
<td>-3.5000</td>
<td>0.000465</td>
</tr>
<tr>
<td>Q2</td>
<td>-1.170</td>
<td>0.241887</td>
<td>-4.4760</td>
<td>0.000008</td>
</tr>
<tr>
<td>Q3</td>
<td>-1.057</td>
<td>0.290578</td>
<td>-1.6150</td>
<td>0.106209</td>
</tr>
<tr>
<td>Q4</td>
<td>-0.943</td>
<td>0.345779</td>
<td>-4.2330</td>
<td>0.000023</td>
</tr>
<tr>
<td>Q5</td>
<td>-1.000</td>
<td>0.317311</td>
<td>-1.7320</td>
<td>0.083265</td>
</tr>
<tr>
<td>Q6</td>
<td>-3.750</td>
<td>0.000177</td>
<td>-4.4090</td>
<td>0.000010</td>
</tr>
<tr>
<td>Q7</td>
<td>-4.666</td>
<td>0.000003</td>
<td>-4.8210</td>
<td>0.000001</td>
</tr>
</tbody>
</table>

Table 3. Detailed Z and p value for all tasks.

The post experiment questionnaires indicate that the touching is considered as the subject’s first choice. This discovery is consistent with the result of the ranked method preference for the mobile AR interaction. Meanwhile, subjects ranked their consideration of significantly influence factors in order of priority, from ease-of-use, accuracy to operational speed.

**Conclusion and Future Work**

In this research, we presented a 3D gesture-based interaction technique for handheld AR applications on a tablet with an external RGB-Depth camera. This allows users to perform 6DOF manipulation of AR virtual objects using their fingers in midair. We evaluated our proposed interaction method by measuring performance (time) and engagement (subjective user feedback). We found that although our method takes longer, it is more natural and enjoyable to use. In the future we will refine the 3D gesture interaction to overcome accuracy limitations and explore scenarios in which 3D gesture interaction is preferred.

**References**


