Impact of Coordinate Systems on 3D Manipulations in Mobile Augmented Reality

Philipp Tiefenbacher
Technische Universität München
philipp.tiefenbacher@tum.de

Steven Wichert
Technische Universität München
stevenwichert@gmail.com

Daniel Merget
Technische Universität München
daniel.merget@tum.de

Gerhard Rigoll
Technische Universität München
rigoll@tum.de

ABSTRACT

Mobile touch PCs allow interactions with virtual objects in augmented reality scenes. Manipulations of 3D objects are a common way of such interactions, which can be performed in three different coordinate systems: the camera-, object- and world coordinate systems. The camera coordinate system changes continuously in augmented reality as it depends on the mobile device's pose. The axis orientations of the world coordinate system are steady, whereas the axes of the object coordinates base on previous manipulations. The selection of a coordinate system therefore influences the 3D transformation's orientation independent from the used manipulation type.
In this paper, we evaluate the impact of the three possible coordinate systems on rotation and on translation of a 3D item in an augmented reality scenario. A study with 36 participants determines the best coordinates for translation and rotation.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*; H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*Interaction styles*

Keywords

interaction; manipulation; augmented reality; mobile; coordinate systems; rotation; translation

1. INTRODUCTION

Mobile touch devices allow easy interaction with complete digital 2D content and mixed reality content. The difficulty of mixed reality content in comparison to pure digital 2D content lies in the extended interaction range of 3D space. The lack of dimension due to the 2D screen raises the question of intuitive manipulation methods. Two essential properties have to be determined for a manipulation in 3D: the axis and the type. In general, the manipulation type is chosen by a certain gesture, which affects the choice of axis, too. For instance, in camera coordinates, sliding a finger to the right should intuitively translate the object to the right.
Henrysson et al. [4] analyze techniques for translation and rotation tasks on mobile devices. The main weakness in their study are the stationary subjects and the imprecise marker tracking. Different positions of the user reveal the specialties of the coordinate systems (CSs), that is why we investigate their influence in a mobile setup with precise tracking in a CAVE.
Multiple works of touch techniques for 3D object manipulation exist [5], though most works consider only stationary touch devices such as tabletops. We integrate a direct manipulation method, which is the primary input modality on smart devices.
We include three rigid body transformations of objects in 3D space: translation (3 DoF), rotation (3 DoF) and uniform scaling (1 DoF). Since the scaling is of a uniform kind, the coordinate axes are not adjusted separately. We investigate their influence only for translation and rotation.
In summary, the main contributions of this paper are:

- we evaluate the impact of coordinate systems on 3D transformations for a tracked mobile device;
- we identify the most suitable coordinate system for rotation and for translation.

In our setup, a change of the device's position alters the view to the 3D content on the mobile display due to active tracking. This is different from approaches with solid touch walls or tabletops. The camera position on stationary devices is fixed and can be manipulated only virtually [3].
Based on the body-centric design space from [17], our scenario can be described as a combination of mid-air gestures fixed in the world and on-device touch gestures relative to the body. The mid-air and on-device gestures control the device's position and the 3D manipulation, respectively.
The user needs to hold the device for the mid-air gestures. That is why the input technique is of a one-handed kind identical to the direct gestures of [16]. Works of interaction concepts on large stationary devices often include both hands. Two hands allow complex gestures, which change rotation and translation of the object simultaneously [15]. We separate rotation and translation because a single hand restricts the number and the ease of gestures. In addition, users in [7, 11] preferred that separation. A menu permits easy selection of rotation, translation and scaling such as in [12]. Obstacles in the real environment may prevent certain camera poses, which may lead to impossible manipulation tasks, when only a 2 DoF rotation or translation is included as in [12].

ICMI '14 November 12 - 16 2014, Istanbul, Turkey

ACM 978-1-4503-2885-2/14/11...$15.00.
http://dx.doi.org/10.1145/2663204.2663234.

Consequently, we include two gestures for rotation and translation (one and two finger) in order to increase integration [9]. A virtual trackball [2] alters rotation on two axes. A two finger rotation around the object's center manipulates the remaining axis. One finger slides the object in the ground plane for the translation, whereas a two-finger slide moves the object in vertical direction of the selected CS.
Ohnishi et al. [13] proposes an alternative to transformations in standard CSs. Here, the 2D-3D mapping occurs in different virtual surfaces. However, the user has to know the correspondence of the input area and the 3D workspace. Martinet et al. [10] propose the Z-technique for the missing degree, which casts a ray into the 3D scene. The ray's direction depends on the touch position and the camera coordinates, thus this technique extends the standard CS with additional rays for interaction.

2. COORDINATE SYSTEMS

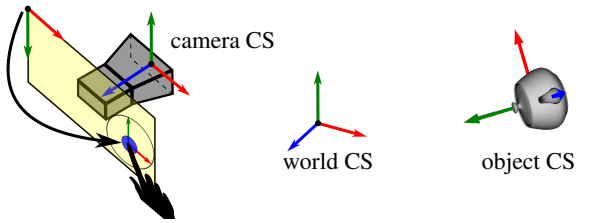

Figure 1: Illustration of the three CSs combined with the mobile device's screen. The direction of a finger move is mapped to one of these CSs, when performing a gesture. The mapping leads to object rotation or translation in the axes directions of the chosen CS. The arrow shows how screen and touch points are linked.

The axes of the camera CS depend on the tablet's current pose, also referred to as the user's view. The user's view is often used as a reference system for mobile AR manipulations [2, 10, 12]. A continuous tracking of the tablet's pose assures correct camera coordinates.
Each object in space has its specific 3D pose. The object CS contains this 3D pose. The object CS's characteristic is the axes' dependencies between translation and rotation. Consequently, a 3D item's rotation changes the axis for the translation also.
The world CS is the only static CS and represents the real world space. Figure 1 depicts all three CSs. Although the CSs are in 3D space, the touch screen of the mobile device is only 2D. Therefore, a mapping between the 3D scene to the 2D display is applied. Only two axes can be manipulated simultaneously on the 2D touch screen. The 3D $\mapsto$ 2D mapping recognizes these two manipulation axes by transforming all three axes with normalized length to the 2D screen. Then the shortest reproduced axis is ignored. The two remaining axes require an association with the horizontal and vertical screen's axis. Therefore, we calculate the absolute slope of the two mapped axes to the tablet PC's horizontal line. The axis with the smaller slope corresponds to the horizontal orientation, the other axis to the screen's vertical direction.

3. USER STUDY

The mobile device of the study is a 11.6" Windows 8 mobile PC. The user study is performed in a CAVE [14]. Lee et al. show in [8] that the completion times of the same task in MR and AR are not significantly different. This finding validates AR studies in our CAVE.

The virtual scene in the CAVE simulates an industrial line. Eight projectors display a realistic 3D scene on the four canvases (left, center, right, bottom). Figure 2 presents a participant during the experiment. Six infrared tracking cameras deliver precise tracking of the targets mounted on the mobile device and the glasses of the subject. The position of the user's head determines the rendered scene in the CAVE. The pose of the mobile device defines the scene on the mobile device's display in parallel. The scene on the tablet

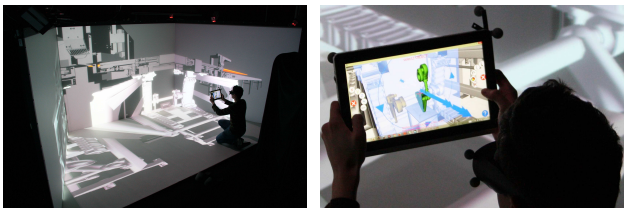

Figure 2: The subject's head and the mobile device are tracked during the study. The mobile device includes additional machine parts, which are not shown in the CAVE.

PC includes the virtual industrial apparatus of the CAVE, but also AR elements which are rendered solely on the mobile device.

3.1 Scenario

Figure 2 presents the main task of the study. The subject steps into the role of an industrial plant fitter, who inserts three 3D machine parts (AR objects) at certain positions. Our goal is to determine differences in the coordinate systems. The objects, therefore, need to be rotated and translated on different axes to match the final pose. For that purpose, the final positions of the three machine parts differ spatially. The biggest distance of the three final poses is around one meter on every axis. For example, one part is close to the ground (0.3 m) and one at a height of 1.3 m.
Besides the manipulable AR parts, colored 3D shadows highlight their final poses in the AR scene. An AR object is fitted successfully as soon as the metrics of rotation, translation and scaling meet defined criteria which are estimated empirically. Thereafter, the subject proceeds by inserting the next AR item into the scene. As the mobile device is tracked during the whole process, the subject is able to move around the AR objects, which have defined positions in world coordinates. Three different AR objects need to be matched in total. The order and initial pose of the inserted 3D machine parts are constant throughout the study. The participants rank their experience in the NASA TLX questionnaire after each completion of the task.

3.2 Distance Metrics

Three different distance metrics are employed to quantify the quality of the AR objects' alignments:

- the rotation metric corresponds to the distance calculation $m_r = ||\mathbf{I} - \mathbf{R}_t \mathbf{R}_i^T||_F$ [6];
- the distance of the current AR object position $\mathbf{T}_i$ to the target $\mathbf{T}_t$ is defined as $m_t = ||\mathbf{T}_t - \mathbf{T}_i||_2^2$;
- the absolute difference characterizes the scaling metric $m_s = ||s_t - s_i||$.

$\mathbf{R}_t$ and $\mathbf{R}_i^T$ are the target and current rotation matrices, respectively. $\mathbf{I}$ is the identity matrix and $||\cdot||_F$ is the Frobenius norm. The rotation metric m_r is zero, when the rotations fit perfectly and $2\sqrt{2}$ in the worst case. We set the thresholds of a successful match to 0.25 for rotation, 0.08 m for translation and 0.15 for scaling.

3.3 Experimental Setup

36 subjects (7 females and 29 males) took part in the study. The average body height was 179.2 *cm*. The participants were aged between 19 and 34 with an average of 23.4. 35 of 36 subjects were accustomed to using touch devices on a regular basis. We split the 36 subjects into three groups. Each group performs its tests in one specific CS for translation. In this way, the comparison between the CSs for translation are obtained due to a between-subject design. Inside each group of 12 subjects, a within-subject design is applied. The within-subject design's main factor is the type of CS for rotation. The order of the CSs for rotation is altered based on the Latin Square design to reduce fatigue and learning effects.

4. RESULTS

The time necessary for the matching task is an evident metric, since the most user-friendly method leads to the fastest task accomplishment. For that reason, we recorded the manipulation times of each transformation type separately. We assume that the rotation times differ as we change the rotations' CSs. Hence, we first exhibit the pure rotation times. Then we compare the translation times between the test groups. Afterwards, we detail the complete manipulation times, which include the durations for rotation, translation and scaling. A subjective analysis of the different CSs is given with the R-TLX scores.

Henceforth, we use the following naming scheme. The capital letters define the kind of transformation. R for rotation and T for translation. The index states the used CS: c = camera, o = object and w = world, and we define the quantity $S = \{c, o, w\}$ with $x \in S$.

4.1 Rotation Time

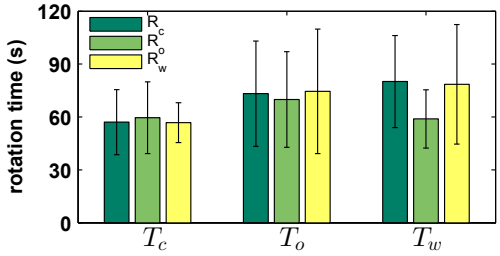

Figure 3: Mean and standard deviation of the rotation time (s).

Each group T_x is analyzed separately. Figure 3 shows the mean and standard deviation of the rotation times. First, the rotation times of the three groups are tested with the Shapiro-Wilk normality test at significance level of .05. The rotation times in T_c and T_o are not distributed normally, thus a Friedman test is applied, which leads to no significant differences. The times in T_w are distributed normally. Consequently, we perform an outlier detection based on [1], which results in the deletion of one entry. A 3x11 uni-variate ANOVA (analysis of variances) yields $F(3,30) = 3.92$, $p = .037 < .05$. Our repeated measure design requires a separate treatment of each subject besides the main factor. The subjects of T_w are significantly faster with R_o (M = 59.52 *s*, SD = 17.18 *s*) than with the two other CSs. The subjects of T_o are also fastest in R_o. In group T_c, R_o is the slowest method, however, this group is in average the fastest of all three groups.

4.2 Translation Time

Again, we perform a Shapiro-Wilk test, which finds that the translation times are not distributed normally. Hence, we deploy the Friedman test. Here, an analysis of the translation times in the same rotation CS but in different T_x indicates no significant differences. Even though the results do not differ significantly, an inspection of Figure 4 indicates a general tendency between each T_x. The translation times in T_c have the lowest mean times independent from the CS for rotation R_x.

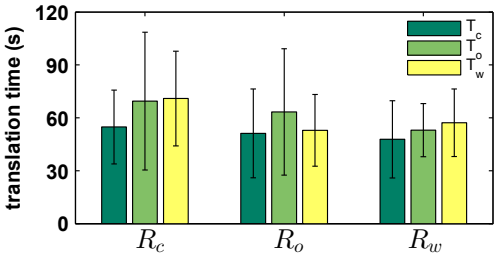

Figure 4: Mean and standard deviation of the transl. time (s).

Conducting the manipulations in T_o yields the worst translation time in R_o. This seems reasonable, because an AR item's rotation leads to different directions of the axes for the translation. The translation times are highest in T_w, when rotating in R_c or R_w.

4.3 Manipulation Time

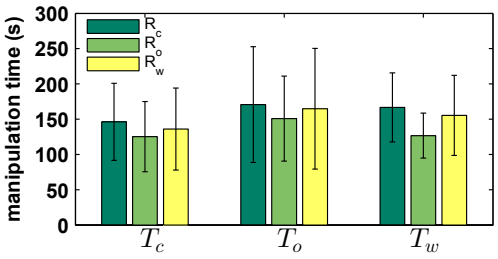

Figure 5: Mean and standard deviation of the manip. time (s).

Figure 5 comprises the complete manipulation times of the 36 subjects. Interestingly, the mean times of all three groups T_x are smallest when rotating the AR items in R_o. Furthermore, an analysis of the manipulation times in T_w shows no normal distribution regarding the Shapiro-Wilk test. A Friedman test results in significant variation of the manipulation time, $\mathcal{X}^2(2,22) = 6.5$, $p = .039 < .05$. Rotating an object in R_c needs the most time in each group T_x.

We compare the manipulation times obtained in the same R_x but in different CSs for translation. In this way, the fastest CS for translation can be determined. In every comparison, the group which positions the object in T_c is the fastest.

4.4 Questionnaires

The NASA-RTLX score indicates the subjective perceived strains, which are caused by mental, physical or temporal demands of the task. The physical workload remains the same throughout the study, because the device and the position of the AR items stays the same. Consequently, the differences in the mean RTLX of Figure 6 highlight mental and temporal demands mostly. The RTLX means are between 47.92 and 56.92 in a total range of 120. Performing the rotation in R_o leads to the smallest RTLX in groups T_c and T_w. In group T_o, R_o is not recommended as stated before. Here, R_w scores best. The scores of R_c and R_w are lowest in T_o, and R_o is lowest in T_w.

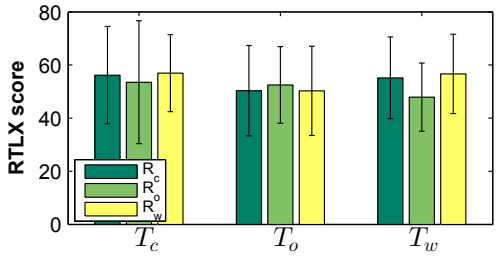

Figure 6: Mean and standard deviation of the raw TLX.

5. CONCLUSION & OUTLOOK

In this work, we detailed possible shortcomings of each CS and described a study for comparing the three possible CSs for translation and rotation separately. We conducted the study with 36 participants in a mixed reality setup, which enables repetitive user-studies since the environmental influences are reduced to a minimum. The participants had to match three different AR items at various spatial positions to ensure different camera poses.

The noticeable variances in the recorded data arise due to learning effects since none of our subjects experienced a mobile AR manipulation task before. Latin Square design, however, ensures similar influence of this effect between the methods.

First, we analyzed the recorded rotation times of the subjects. The findings suggest that the favorite CS for rotation is the object CS. Groups T_o and T_w accomplish the rotation fastest, when working in R_o. Additionally, the manipulation times are lowest in each group when using R_o. The different CSs for translation were split into three separate groups. Each group performed the task in a single CS for translation. We observed that T_c is the best one in regard to the translation time and the complete manipulation time.

It is interesting to note, that the subjects are faster with a static CS for rotation, but the user's view dependent CS is better for the translation. The use of R_o decouples the user's position from the rotation task. This way, the user can concentrate on the rotation task regardless of the position. We believe that this reduces the difficulty, because the user does not have to find the best position for matching the object. Instead the user can rely on the knowledge about the axis orientations.

This idea manifests further when considering the subjective RTLX results. Translation in T_o is the least straining CS, besides the combination R_oT_o as discussed in Section 4.2.

Taken together, these finding suggest that it is easiest for the user to perform the transformation in the object CS. However, with slightly more effort in the camera CS, the users are faster at translating. Conducting the rotation in camera CS does not improve the manipulation time as it is the case for the translation.

We are of the opinion that the AR item's rotation requires more spatial imagination than a translation, which complicates the rotation task for some users. Our results are in good agreement with Henrysson et al. [4], who concluded that performing a translation based on the device's position is better than with the static keypad technique but the static keypad is better for a rotation task.

Consequently, we recommend the object CS for rotation and the camera CS for translation in a mobile AR scenario.

Further studies, which take other input concepts such as HOMER-S [12] into account, have to be undertaken to generalize our findings. Further work needs to be done to estimate the learning effects on the three CSs by conducting a study with multiple iterations in the same CS.

6. ACKNOWLEDGEMENTS

The research leading to these inventions has received funding from the European Union Seventh Framework Programm (FP7/2007-2013) under grant agreement n° 284573.

7. REFERENCES

[1] V. Barnett and T. Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[2] D. Fiorella, A. Sanna, and F. Lamberti. Multi-touch User Interface Evaluation for 3D Object Manipulation on Mobile Devices. *Multimodal User Interfaces*, 4(1):3–10, 2010.

[3] M. Hancock, T. ten Cate, and S. Carpendale. Sticky tools: full 6DOF force-based interaction for multi-touch tables. In *Proc. ITS*, pages 133–140. ACM, 2009.

[4] A. Henrysson, J. Marshall, and M. Billinghurst. Experiments in 3D interaction for mobile phone AR. In *Proc. GRAPHITE*, pages 187–194. ACM, 2007.

[5] U. Hinrichs and S. Carpendale. Gestures in the wild: studying multi-touch gesture sequences on interactive tabletop exhibits. In *Proc. CHI*, pages 3023–3032. ACM, 2011.

[6] D. Q. Huynh. Metrics for 3D Rotations: Comparison and Analysis. *Mathematical Imaging and Vision*, 35(2):155–164, 2009.

[7] K. Kin, T. Miller, B. Bollensdorff, T. DeRose, B. Hartmann, and M. Agrawala. Eden: a professional multitouch tool for constructing virtual organic environments. In *Proc. SIGCHI*, pages 1343–1352. ACM, 2011.

[8] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman. The Effects of Visual Realism on Search Tasks in Mixed Reality Simulation. *Visualization and Computer Graphics*, 19(4):547–556, 2013.

[9] J. Liu, O. K.-C. Au, H. Fu, and C.-L. Tai. Two-Finger Gestures for 6DOF Manipulation of 3D Objects. *Computer Graphics Forum*, 31(7):2047–2055, 2012.

[10] A. Martinet, G. Casiez, and L. Grisoni. The design and evaluation of 3D positioning techniques for multi-touch displays. In *Proc. 3DUI*, pages 115–118. IEEE, 2010.

[11] A. Martinet, G. Casiez, and L. Grisoni. Integrality and separability of multitouch interaction techniques in 3d manipulation tasks. *Visualization and Computer Graphics*, 18(3):369–380, 2012.

[12] A. Mossel, B. Venditti, and H. Kaufmann. 3DTouch and HOMER-S: Intuitive Manipulation Techniques for One-Handed Handheld Augmented Reality. In *Proc. VRIC*, pages 12:1–12:10. ACM, 2013.

[13] T. Ohnishi, N. Katzakis, K. Kiyokawa, and H. Takemura. Virtual interaction surface: Decoupling of interaction and view dimensions for flexible indirect 3D interaction. In *Proc. 3DUI*, pages 113–116. IEEE, 2012.

[14] E. Ragan, C. Wilkes, D. A. Bowman, and T. Höllerer. Simulation of augmented reality systems in purely virtual environments. In *Proc. VR*, pages 287–288. IEEE, 2009.

[15] J. L. Reisman, P. L. Davidson, and J. Y. Han. A screen-space formulation for 2D and 3D direct manipulation. In *Proc. UIST*, pages 69–78. ACM, 2009.

[16] P. Tiefenbacher, A. Pflaum, and G. Rigoll. Touch Gestures for Improved 3D Object Manipulation in Mobile Augmented Reality. In *Proc. ISMAR*. IEEE, 2014.

[17] J. Wagner, M. Nancel, S. G. Gustafson, S. Huot, and W. E. Mackay. Body-centric Design Space for Multi-surface Interaction. In *Proc. CHI*, pages 1299–1308. ACM, 2013.