Multimodal Interaction Concepts for Mobile Augmented Reality Applications

Wolfgang Hürst and Casper van Wezel

Utrecht University, PO Box 80.089, 3508 TB Utrecht, The Netherlands
huerst@cs.uu.nl, cawezel@students.cs.uu.nl

Abstract. Augmented reality on mobile phones – i.e. applications where users look at the live image of the device's video camera and the scene that they see is enriched by 3D virtual objects – provides great potential in areas such as cultural heritage, entertainment, and tourism. However, current interaction concepts are often limited to pure 2D pointing and clicking on the device's screen. This paper explores different interaction approaches that rely on multimodal sensor input and aim at providing a richer, more complex, and engaging interaction experience. We present a user study that investigates the usefulness of our approaches, verify their usability, and identify limitations as well as possibilities for interaction development for mobile augmented reality.

Keywords: Mobile augmented reality, multi-sensor input, interaction design.

1 Introduction

In this paper, we explore new interaction metaphors for mobile augmented reality (***mobile AR***), i.e. applications where users look at the live image of the video camera on their mobile phone and the scene that they see (i.e. the reality) is enriched by integrated 3-dimensional virtual objects (i.e. an augmented reality; cf. Fig. 1). Even if virtual objects are registered in 3D and updated in real-time, current interaction concepts are often limited to pure 2D pointing and clicking via the device's touch screen. However, in order to explore the tremendous potential of augmented reality on mobile phones for areas such as cultural heritage, entertainment, or tourism, users need to create, access, modify, and annotated virtual objects and their relations to the real world by manipulations in 3D.

In addition to common problems with interaction on mobile phones such as small screen estate, interaction design for mobile AR is particularly difficult due to several characteristics. First of all, since we are interacting with the (augmented) reality, interface design is somehow limited. Normally, an appropriate interface design can often deal with mobile interaction problems. For example, one can limit the number of buttons and increase their sizes so they can easily be hit despite small screen sizes. In mobile AR however, placement, style, and size of virtual objects is dictated by the video image and the virtual augmentation, resulting for example in sizes of objects that are hard to select (cf. Fig. 2, top). In addition, interfaces of normal applications can be operated while holding the device in a stable and comfortable position,

K.-T. Lee et al. (Eds.): MMM 2011, Part II, LNCS 6524, pp. 157–167, 2011.

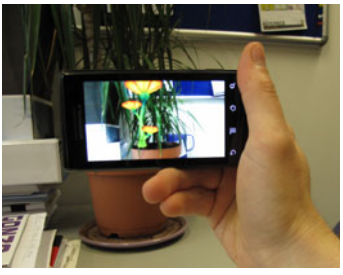
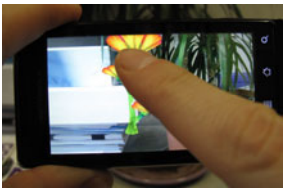


Fig. 1. Mobile Augmented Reality enhances reality (represented by the live video stream on your mobile) with 3D virtual objects (superimposed at a position in the video that corresponds to a related spot in the real world). In this example, a virtual flower grows in the real plant pot.

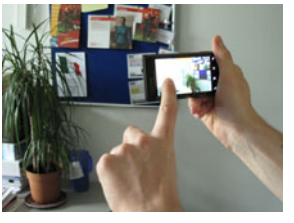
whereas mobile AR applications require users to point their phone to a certain scene, often resulting in uncomfortable and unstable positions of the phone (especially when touching the screen at the same time; cf. Fig. 2, center). Being forced to point the device to a specific position in space also limits the possibility to interact via tilting (which has become one of the most commonly used interaction metaphors for controlling virtual reality data on mobiles e.g. for mobile gaming). Finally, input signals resulting from touch screen interactions are 2-dimensional, whereas augmented reality requires us to manipulate objects in 3D (cf. Fig. 2, bottom).

The goal of this paper is to explore alternatives to touch screen based interaction in mobile AR that take advantage of multimodal information delivered by the sensors integrated in modern smart phones, i.e. camera, accelerometer, and compass. In particular, the accelerometer combined with the compass can be used to get the orientation of the phone. In fact, these sensors are used to create the augmented reality in the first place by specifying where and how virtual 3D objects are registered in the scene of the real world. We present an approach that uses this information also to select and manipulate virtual 3D objects. We compare it with touch screen interaction and a third approach that analyzes the camera image. In the latter case, the tip of your finger is tracked when it is moved in front of the camera and gestural input is used for object manipulation. Using gestural input is currently a hot trend in human computer interaction as illustrated by projects such as Microsoft's Natal (now Kinect) [1] and MIT's SixthSense [2]. Gesture-based interaction on mobile phones has also been explored by both industry [3] and academia [4] but in relation to different applications than augmented reality and utilizing a user-facing camera. Work in mobile AR that takes advantage of the front facing camera include analysis of hand-drawn shapes [5] and 3D sketches [6] as well as various marker-based approaches such as tangible interfaces [7]. However, most of these have been applied in indoor environments under rather restricted and controlled conditions, and often rely on additional tools such as the utilization of a pen or markers. [8] is an example for work in traditional augmented reality (e.g. using head mounted displays) that utilizes hand gestures. However, we are unaware of any work applying this concept to augmented reality on mobile phones. The purpose of the study presented in this paper is to verify if this concept is also suitable in a mobile context. In the following, we describe the three interaction concepts and tasks we want to evaluate (section 2), present our user study (section 3) and results (section 4), and conclude with a discussion about the consequences and our resulting future work on interaction design for mobile AR (section 5).

Interface design. A good interface design allows us to deal with many interaction problems, e.g. by making buttons big enough or enlarging them during interaction for better visibility (left). In augmented reality, size and position of the objects is dictated by reality and thus objects might be too small to easily select or manipulate them (right).



Holding the device. Whereas normally one can hold the device in a stable and comfortable position during interaction (left), augmented reality requires us to point the phone to a certain spot in the real world, resulting in an unstable position esp. during interaction (right).



2D vs. 3D. Touch screen based interaction only delivers 2-dimensional data, hence making it difficult to control virtual objects in the 3D world (e.g. put a flower into a plant pot; right).

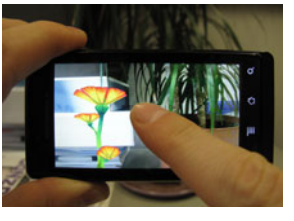


Fig. 2. Potential problems with mobile augmented reality interaction

2 Interaction Concepts and Tasks

Our goal is to compare standard touch screen based interaction with two different interaction concepts for mobile AR: one that depends on how the device is moved (utilizing accelerometer and compass sensors) and one that tracks the user's finger in front of the camera (utilizing the camera sensor). Our overall aim is to target more complex operations with virtual objects than pure clicking. In the ideal case, a system should support all kinds of object manipulations in 3D, i.e. selection, translation, scaling, and rotation. However, for an initial study and in order to be able to better compare it to touch screen based interaction (which per default only delivers 2-dimensional data), we restrict ourselves here to the three ***tasks*** of *selecting virtual objects*, *selecting entries in context menus*, and *translation of 3D objects in 2D* (i.e. left/right and up/down). In order to better investigate the actual interaction experience and eliminate noise from hand and finger tracking, we use a marker that is attached to the finger that will be tracked (cf. below).

For the standard ***touch screen based interaction***, the three tasks have been implemented in the following way: selecting an object is achieved by simply clicking on it on the touch screen (cf. Fig. 3, left). This selection evokes a context menu that was implemented in a pie menu style [9] which has proven to be more effective for pen and finger based interaction (compared to list-like menus as commonly used for mouse or touchpad based interaction; cf. Fig. 3, right). One of these menu entries puts

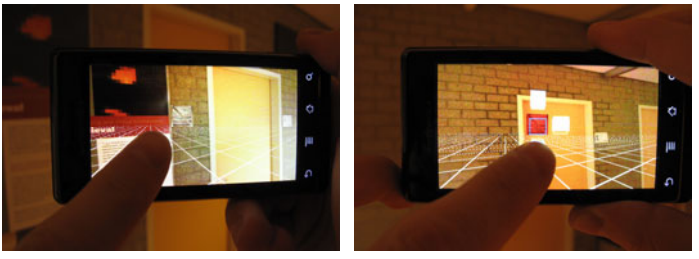


Fig. 3. Touch screen interaction: select an object (left) or an entry from a context menu that pops up around the object after selection (right)

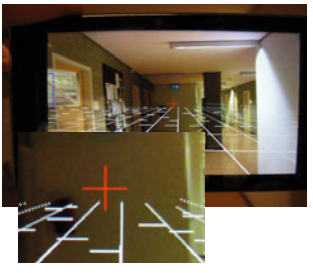
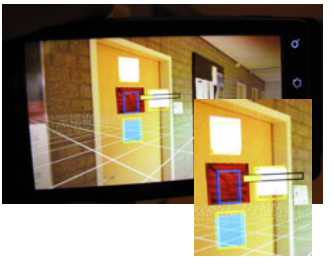


Fig. 4. Device based interaction: selection by pointing a reticule (left) to the target till the related bar is filled (right)



the object in "translation mode" in which a user can move an object around by clicking on it and dragging it across the screen. If the device is moved without dragging the object, it stays at its position with respect to the real world (represented by the live video stream). Leaving translation mode by clicking at a related icon fixes the object at its new final position in the real world.

In terms of usability, we expect this approach to be simple and intuitive because it conforms to standard touch screen based interaction. In case of menu selection, we can also expect it to be reliable and accurate because we are in full control of the interface design (e.g. we can make the menu entries large enough and placed as far apart of each other so they can be hit easily with your finger). We expect some accuracy problems though when selecting virtual object, especially if they are rather small, very close to each other, or overlapping because they are placed behind each other in the 3D world. In addition, we expect that users might feel uncomfortable when interacting with the touch screen while they have to hold the device upright and targeted towards a specific position in the real world (cf. Fig. 2, center). This might be particularly true in the translation task for situations where the object has to be moved to a position that is not shown in the original video image (e.g. if users want to place it somewhere behind them).

Our second interaction concept uses the ***position and orientation of the device*** (defined by the data delivered from the integrated accelerometer and compass) for interaction. In this case, a reticule is visualized in the center of the screen (cf. Fig. 4, left) and used for selection and translation. Holding it over an object for a certain amount of time (1.25 sec in our implementation) selects the object and evokes the pie menu. In order to avoid accidental selection of random objects, a progress bar is shown above the object to illustrate the time till it is selected. Menu selection works in the same way by moving the reticule over one of the entries and holding it till the bar is filled (cf. Fig. 4, right). In translation mode, the object sticks to the reticule in

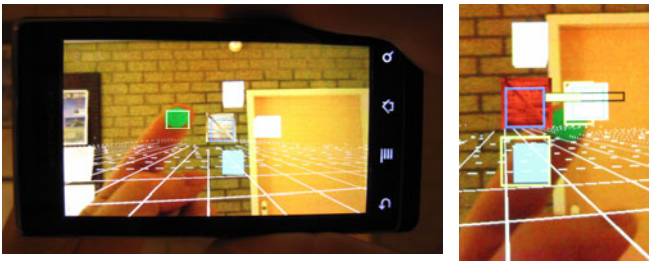


Fig. 5. Interaction by using a green marker for tracking that is attached to the tip of the finger

the center of the screen while the device is moved around. It can be placed at a target position by clicking anywhere on the touch screen. This action also forces the system to leave translation mode and go back to normal interaction.

Compared to touch screen interaction, we expect this "device based" interaction to take longer when selecting objects and menu entries because users can not directly select them but have to wait till the progress bar is filled. In terms of accuracy, it might allow for a more precise selection because the reticule could be pointed more accurately at a rather small target than your finger. However, holding the device at one position over a longer period of time (even if it's just 1.25 sec) might prove to be critical especially when the device is hold straight up into the air. Translation with this approach seems intuitive and might be easier to handle because people just have to move the device (in contrast to the touch screen where they have to move the device and drag the object at the same time). However, placing the object at a target position by clicking on the touch screen might introduce some inaccuracy because we can expect that the device shakes a little when being touched while held up straight in the air. This is also true for the touch screen based interaction, but might be more critical here because for touch screen interaction the finger already rests on the screen. Hence, we just have to it release it whereas here we have to explicitly click the icon (i.e. perform a click and release action).

Touch screen based interaction seems intuitive because it conforms to regular smart phone interaction with almost all common applications (including most current commercial mobile AR programs). However, it only allows to remotely controlling the 3-dimensional augmented reality via 2D input on the touch screen. If we track the users' index finger when their hand is moved in front of the device (i.e. when it appears in the live video on the screen), we can realize a ***finger based interaction*** where the tip of the finger can be used to directly interact with objects, i.e. select and manipulate them. In the ideal case, we can track the finger in all three dimensions and thus enable full manipulation of objects in 3D. However, since 3D tracking with a single camera is difficult and noisy (especially on a mobile phone with a moving camera and relatively low processing power) we restrict ourselves to 2D interactions in the study presented in this paper. In order to avoid influences of noisy input from the tracking algorithm, we also decided to use a robust marker based tracking approach where users attach a small sticker to the tip of their index finger (cf. Fig. 5). Object selection is done by "touching" an object (i.e. holding the finger at the position in the real world where the virtual object is displayed on the screen) till an associated

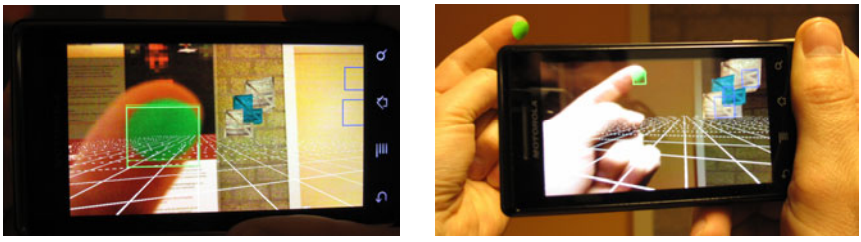


Fig. 6. Holding the finger too close to the camera makes it impossible to select small objects (left). Moving your hand away from the camera decreases the size of the finger in the image but can result in an uncomfortable position because you have to stretch out your arm (right).

progress bar is filled. Menu entries can be selected in a similar fashion. In translation mode, objects can be moved by "pushing" them. For example, an object is moved to the right by approaching it with the finger from the left side and pushing it rightwards. Clicking anywhere on the touch screen places the object at its final position and leaves translation mode.

Gesture based interaction using finger or hand tracking can be a very powerful way for human computer interaction in a lot of situations. However, in relation to mobile AR, there are many shortcomings and potential problems, most importantly due to the limited range covered by the camera. For example, the finger becomes very large in the image when being close to the camera, making it unsuitable for selection of small or overlapping objects. Moving it further away should increase possibilities for interaction, but might result in uncomfortable positions where the arm has to be stretched out quite far in order to create a smaller image of the finger on the screen (cf. Fig. 6). In addition, moving an object by pushing it from the side might turn out to be difficult depending on which hand is used (e.g. pushing it from the right using the left hand) and either result in an awkward hand position or even force people to switch the hand in which they hold the device.

We implemented all three interaction concepts on a Motorola Droid/Milestone phone with Android OS version 2.1. In the next chapter we present a user study to verify their intuitive advantages and disadvantages discussed above.

3 User Study

We evaluated the interface concepts described in the previous section in a user study with 18 participants (12 male and 6 female, 5 users at ages 15-20 years, 8 at ages 21-30, 1 at ages 31-40 and 41-50, and 3 at ages 51-52). For the finger tracking, users were free in where to place the marker on their finger tip. Only one user placed it on his nail. All others used it on the inner side of their finger as shown in Figures 5 and 6. Eleven participants held the device in their right hand and used the left hand for interaction. For the other seven it was just the other way around. No differences could be observed in the evaluation related to marker placement or hand usage.

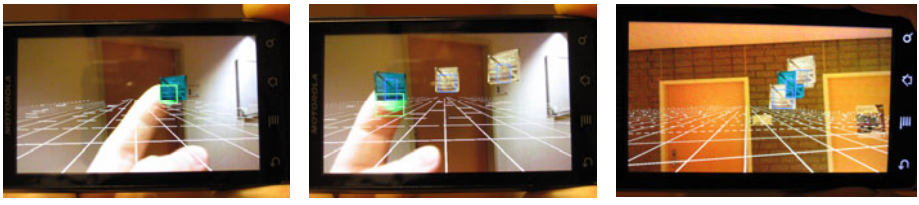


Fig. 7. Object selection task: test case (single object, left), easy (multiple non-overlapping objects, center), and hard (multiple overlapping objects, right)

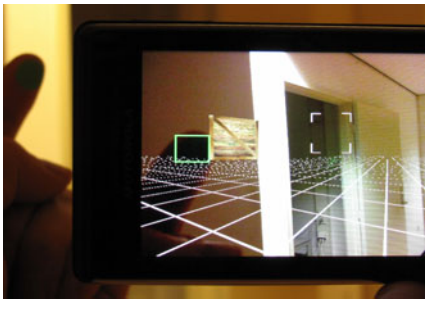


Fig. 8. Translation task: move object to target (white square on the right)

A within-group study was used, i.e. each participant tested each interface (subsequently called ***touch screen***, ***device***, and ***finger***) and task (subsequently called ***object selection***, ***menu selection***, and ***translation***). Interfaces were presented in different order to the participants to exclude potential learning effects. For each user, tasks were done in the following order: object selection, then menu selection, then translation, because this would also be the natural order in a real usage case. For each task, there was one introduction test in which the interaction method was explained to the subject, one practice test in which the subject could try it out, and finally three "real" tests (four in case of the translation task) that were used in the evaluation. Subjects were aware that the first two tests were not part of the actual experiment and were told to perform the other tests as fast and accurate as possible. The three tests used in the object selection task can be classified as easy (objects were placed far away from each other), medium (objects were closer together), and hard (objects overlapped; cf. Fig. 7). In the menu selection task, the menu contained three entries and users had to select one of the entries on top, at the bottom, and to the right in each of the three tests (cf. Fig. 3, 4, and 5, right). In the translation task, subjects had to move an object to an indicated target position (cf. Fig. 8). The view in one image covered a range of 72.5°. In two of the four tests, the target position was placed within the same window as the object that had to be moved (at an angle of 35° between target and initial object position to the left and right, respectively). In the other two tests, the target was outside of the initial view but users were told in which direction they had to look to find it. It was placed at an angel of 130° between target and object, to the left in one case, and to the right in the other one. The order of tests was randomized for each participant to avoid any order-related influences on the results.

For the evaluation, we logged the time it took to complete the task, success or failure, and all input data (i.e. the sensor data delivered from the accelerometer, compass, and the marker tracker). Since entertainment and leisure applications play an important role in mobile computing, we were not only interested in pure accuracy and performance, but also in issues such as fun, engagement, and individual preference.

Hence, users had to fill out a related questionnaire and were interviewed and asked about their feedback and further comments at the end of the evaluation.

4 Results

Figure 9 shows how long it took the subjects to complete the tests averaged over all users for each task (Fig. 9, top left) and for individual tests within one task (Fig. 9, top right and bottom). The averages in Figure 9, top left show that touch screen is the fastest approach for both selection tasks. This observation conforms to our expectations mentioned in the previous section. For device and finger, selection times are longer but still seem reasonable.

Looking at the individual tests used in each of these tasks, times in the menu selection task seem to be independent of the position of the selected menu entry (cf. Fig. 9, bottom left). Almost all tests were performed correctly: only three mistakes happened over all with the device approach and one with the finger approach.

In case of the object selection task (cf. Fig. 9, top right), touch screen interaction again performed fastest and there were no differences for the different levels of difficulty of the tasks. However, there was one mistake among all easy tests and in five of the hard tests the wrong object was selected thus confirming our assumption that interaction via touch screen will be critical in terms of accuracy for small or close objects. Finger interaction worked more accurate with only two mistakes in the hard test. However, this came at the price of a large increase of selection time. Looking into the data we realized that this was mostly due to subjects holding the finger relatively close to the camera resulting in a large marker that made it difficult to select an individual object that was partly overlapped by others. Once the users moved their hand further away from the camera, selection worked well as indicated by the low amounts of errors. For the device approach, there was a relatively large number of errors for the hard test, but looking into the data we realized that this was only due to a mistake that we made in the setup of the experiment: in all six cases, the reticule was already placed over an object when the test started and the subjects did not move the device away from it fast enough to avoid accidental selection. If we eliminate these users from the test set, the time illustrated for the device approach in the hard test illustrated in Figure 9, top right increases from 10,618 msec to 14,741 msec which is still in about the same range as the time used for the tests with easy and medium levels of difficulty. Since all tests in which the initialization problem did not happen have been solved correctly, we can conclude that being forced to point the device to a particular position over a longer period of time did not result in accuracy problems as we suspected.

In the translation task, we see an expected increase in time in case of the finger and touch screen approach if the target position is outside of the original viewing window (cf. Fig. 9, bottom right). In contrast to this, there are only small increases in the time it took to solve the tasks when the device approach is used. In order to verify the quality of the results for the translation task, we calculated the difference between the center of the target and the actual placement of the object by the participants in the tests. Figure 10 illustrates these differences averaged over all users. It can be seen that

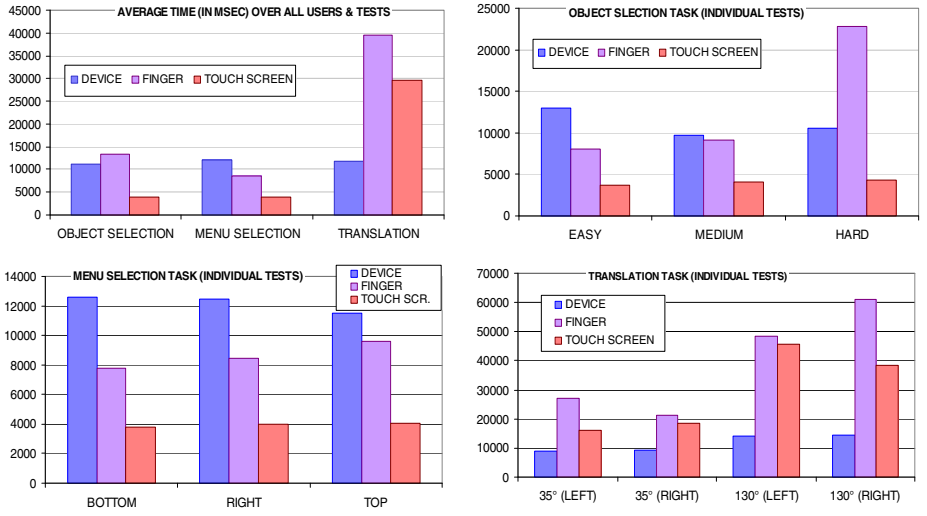


Fig. 9. Time it took to solve the tasks (in msec, averaged over all users)

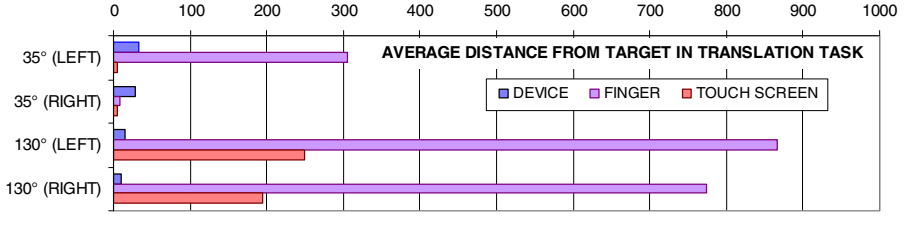


Fig. 10. Correctness of translation task (squared Manhattan distance in screen units)

the device approach was not only the fastest but also very accurate, especially in the more difficult cases with an angel of 130° between target and initial object position. Finger based interaction on the other hand seems very inaccurate. A closer look into the data reveals that the reason for the high difference between target position and actual object placement are actually many situations in which the participants accidentally hit the touch screen and thus placed the object at a random position before reaching the actual target. This illustrates a basic problem with finger based interaction, i.e. that users have to concentrate on two issues: moving the device and "pushing" the object with their finger at the same time. In case of device based interaction on the other hand, users could comfortably take the device in both hands while moving around thus resulting in no erroneous input and a high accuracy as well as fast solution time. A closer look into the data also revealed that the lower performance in case of the touch based interaction are mostly due to one outlier who did extremely bad (most likely also due to an accidental input) but otherwise they are at about the same level of accuracy as the device approach (but of course at the price of a worse performance time; cf. Fig. 9).

Based on this formal analysis of the data, we can conclude that touch based interaction seems appropriate for selection tasks. For translation tasks, the device based approach seems to be more suitable, whereas finger based interaction seems less useful in general. However, especially in entertainment and leisure applications, speed and accuracy are not the only relevant issues, but fun and engagement can be equally important. In fact, for gaming, mastering an inaccurate interface might actually be the whole purpose of the game (think of balancing a marble through a maze by tilting your phone appropriately – an interaction task that can be fun but is by no means easy to handle). Hence, we asked the participants to rank the interaction concepts based on performance, fun, and both. Results are summarized in Table 1. Rankings for performance clearly reflect the results illustrated in Figures 9 and 10 with touch based interaction ranked highest by eleven participants, and device ranked first by six. Only one user listed finger based interaction as top choice for performance. However, when it comes to fun and engagement, the vast majority of subjects ranked it as their top choice, whereas device and touch were ranked first only four and one time, respectively. Consequently, rankings are more diverse when users were asked to consider both issues. No clear "winner" can be identified here, and in fact, many users commented that it depends on the task: touch and device based interaction are more appropriate for applications requiring accurate placement and control, whereas finger based interaction was considered as very suitable for gaming applications. Typical comments about the finger based approach characterized it as "fun" and "the coolest" of all three. However, its handling was also criticized as summarized by one user who described it as "challenging but fun". When asked about the usefulness of the three approaches for the three tasks of the evaluation, rankings clearly reflect the objective data discussed above. Table 2 shows that all participants ranked touch based interaction first for both selection tasks, but the vast majority voted for the device approach in case of translation. It is somehow surprising though that despite its low performance, finger based interaction was ranked second by eight, seven, and nine users, respectively, in each of the three tasks – another indication that people enjoyed and liked the approach although it is much harder to handle than the other two.

Table 1. Times how often an interface was ranked first, second, and third with respect to performance versus fun and engagement versus both (T = touch screen, D = device, F = finger)

	PERFORMANCE			FUN			PERF. & FUN		
	T	D	F	T	D	F	T	D	F
Ranked 1st	11	6	1	1	4	13	8	3	7
Ranked 2nd	3	6	9	6	10	2	6	8	4
Ranked 3rd	4	6	8	11	4	3	4	7	7

Table 2. Times how often an interface was ranked first, second, and third with respect to the individual tasks (T = touch screen, D = device, F = finger)

	OBJECT SELECT.			MENU SELECT.			TRANSLATION		
	T	D	F	T	D	F	T	D	F
Ranked 1st	18	-	-	18	-	-	3	15	-
Ranked 2nd	-	10	8	-	11	7	6	3	9
Ranked 3rd	-	8	10	-	7	11	9	-	9

5 Conclusion

Interacting with augmented realities on mobile devices is a comprehensive experience covering many ranges, characteristics, and situations. Hence, we can not assume to find a "one size fits all" solution for a good interface design, but most likely have to provide different means of interaction depending on the application, the goal of a particular interaction, and the context and preference of the user. Our study suggests, that a combination of touch screen based interaction (which achieved the best results in the shortest time in both selection tasks and was rated highest by users in terms of performance) with device dependent input (which achieved the best results in the shortest time in the translation task and was highly ranked by the users for this purpose) is a promising approach for serious applications that require exact positioning and accurate control over the content. On the other hand, interaction via finger tracking (which had low performance values but was highly rated and appreciated by the users in terms of fun, engagement, and entertainment) seems to be a promising approach for mobile gaming and other leisure applications.

Obviously, our study was only intended as a first step in the direction of better, more advanced interfaces for mobile AR. In relation to serious applications, our future work aims at further investigating touch and device based navigation, especially in relation to other tasks, such as rotation and scaling of objects. In relation to leisure games, our goal is to further investigate interaction via finger and hand tracking, especially in relation to "real" 3D interaction where moving the finger in 3D space can be used to manipulate objects in 3D, for example by not only pushing them left/right and up/down but also forward and backward.

References

1. Xbox Kinect (formerly known as Microsoft's project Natal), http://www.xbox.com/en-US/kinect (last accessed 10/15/10)
2. Mistry, P., Maes, P.: SixthSense: a wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches (2009)
3. EyeSight's Touch Free Interface Technology Software, http://www.eyesight-tech.com/technology/ (last accessed 10/15/10)
4. Niikura, T., Hirobe, Y., Cassinelli, A., Watanabe, Y., Komuro, T., Ishikawa, M.: In-air typing interface for mobile devices with vibration feedback. In: ACM SIGGRAPH 2010 Emerging Technologies (2010)
5. Hagbi, N., Bergig, O., El-Sana, J., Billinghurst, M.: Shape recognition and pose estimation for mobile augmented reality. In: Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality (2009)
6. Bergig, O., Hagbi, N., El-Sana, J., Billinghurst, M.: In-place 3D sketching for authoring and augmenting mechanical systems. In: Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality (2009)
7. Billinghurst, M., Kato, H., Myojin, S.: Advanced Interaction Techniques for Augmented Reality Applications. In: Proceedings of the 3rd international Conference on Virtual and Mixed Realit, Part of HCI International 2009 (2009)
8. Lee, M., Green, R., Billinghurst, M.: 3D natural hand interaction for AR applications. In: Proceedings of Image and Vision Computing, New Zealand (2008)
9. Callahan, J., Hopkins, D., Weiser, M., Shneiderman, B.: An empirical comparison of pie vs. linear menus. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (1988)