Who's That Girl? Handheld Augmented Reality for Printed Photo Books

Niels Henze and Susanne Boll

University of Oldenburg
Escherweg 2, 26121 Oldenburg, Germany
{niels.henze,susanne.boll}@uni-oldenburg.de

Abstract. Augmented reality on mobile phones has recently made major progress. Lightweight, markerless object recognition and tracking makes handheld Augmented Reality feasible for new application domains. As this field is technology driven the interface design has mostly been neglected. In this paper we investigate visualization techniques for augmenting printed documents using handheld Augmented Reality. We selected the augmentation of printed photo books as our application domain because photo books are enduring artefacts that often have online galleries containing further information as digital counterpart. Based on an initial study, we designed two augmentations and three techniques to select regions in photos. In an experiment, we compare an augmentation that is aligned to the phone's display with an augmentation aligned to the physical object. We conclude that an object aligned presentation is more usable. For selecting regions we show that participants are more satisfied using simple touch input compared to Augmented Reality based input techniques.

Keywords: augmented reality, mobile phone, photo sharing, mobile interaction, image analysis, photo book.

1 Introduction

The paperless office was predicted more than a quarter century ago. Despite the availability of desktop computers, notebooks and smartphones everywhere at any time this revolution has not happened yet [25]. Paper documents have qualities that current digital device fail to offer: paper has an amazing resolution, does not need a power supply, and remains accessible for centuries; furthermore, paper is cheap and one can leave it lying around without much concern about thievery or environmental influences. The digital revolution, however, happened in the last decades. While the digital world is increasingly getting dynamic and interactive - paper remains static. Once printed or written the content remains. Today, paper documents are usually produced from digital content and printing is the end of the production chain. Related digital media is often available at production time but is not used for particular prints. Digital content might also evolve, get annotated, iterated and improved but the once printed document provides no access to these changes. What is missing is a way to unify the advantages of printed documents with the flexibility and richness of online content.

P. Campos et al. (Eds.): INTERACT 2011, Part III, LNCS 6948, pp. 134–151, 2011.
© IFIP International Federation for Information Processing 2011

Using mobile devices in conjunction with physical objects has been proposed to unify the advantages of paper documents with the interactivity, personalization, and real-time features provided by digital media [7]. Augmented Reality (AR) looks as a natural option to fuse digital media with paper documents [14]. As traditional AR has severe technical limitation, such as the need for AR goggles, handheld AR recently received a great share of attention in the AR domain [26]. Smartphones are used to present the phone's camera image on the screen augmented with additional content that is aligned to the recorded physical world. Recent technical development has made major progress beyond the use of artificial markers towards augmenting diverse types and objects. Handheld AR research is currently dominated by technical development. New algorithms make handheld AR feasible for more and more different types of object, incremental improvement increases the accuracy, and combining different techniques greatly improved the performance [26]. The interface design, however, has been mostly neglected so far. The HCI community only scratched the surface of designing handheld AR interfaces.

The application domain we selected to investigate the interface design of handheld AR systems is the interaction with printed photo books. Printed photo books are highly emotional enduring artefacts, handcrafted by users to preserve memories and often have interactive online galleries as natural counterparts. To create a connection between the physical and the digital, we try to enable users accessing the same information available in photo communities' galleries using physical photo books.

In this paper, we use photo books to investigate two fundamental aspects of handheld AR for printed documents: How to visualize annotations and how to select regions to create annotations. In the following Section we discuss related research. Afterwards, we motivate the augmentation of printed photo books in Section 3 and we present an initial study to collect design approaches in Section 4. The visualization techniques and selection techniques are designed in Section 5. Section 6 describes the design and the results of the experiment that compares the designed approaches. We close the paper with conclusions and an outlook to future work.

2 Related Work

Several approaches have been developed to create the link between physical objects and digital content. Fitzmaurice was one of the first who predicted the use of mobile devices for interaction with physical objects by simply pointing at them. He described for instance an application with which the user can point at locations on a physical map with a mobile device to get additional information [7]. In recent years more systems emerged that provide information related to physical objects using mobile phones. Common implementation mark objects either with visual markers [21] or with electronic tags [28]. Barcodes can be seen as the first implementation of visual markers back in 1948. Since special barcode readers are needed to read 1D barcodes 2D barcodes such as QR-codes [15] have been developed. QR-codes can be read using recent mobile phones and a few mobile phones also have an integrated RFID reader to read electronic tags. However, not all physical objects are suitable for markers. Sights and buildings are simply too large or out of range to be reasonably equipped with either type of markers. It is also questionable if objects, whose visual

appeal is important, such as printed photo books, can in general be equipped with visual markers. Not only because the markers require visual space but also because visual markers affect the design of the object. Electronic marker, however, lead to ambiguity for applications with a high object density. Furthermore all markers restrict the interaction radius in a specific way. In particular, electronic markers do not to provide the relative location of the object.

Another approach to create the link between physical objects and digital content is using content-based object recognition. Davies et al. investigated user reaction to the use of a digital image capture and recognition system for mobile phone to access information about sights [5]. With their system users can take a photo of a sight with a mobile phone and receives a related description. Davies et al. found that half of the users want to use such a system even when "this is a more complex, lengthy and error-prone process than traditional solutions". Pielot et al. compared the use of object recognition and typing an URL to access information about posters [20] with a mobile phone. They showed that users are faster using object recognition and also prefer this interaction technique. In our own work, we showed that even novice users can use this interaction to get information about printed photos [10].

Handheld AR (also called Magic Lens) is the adaptation of AR for mobile phones. It extends the concept of using discrete photos by providing immediate feedback using an augmentation. Prototypes that augment different types of physical objects, such as text-heavy conference proceedings [14], real estate guides [6], and sights [1] have been developed. Different groups (e.g. [23, 18, 12]) developed prototypes to augment paper maps and conducted according user studies. Rohs et al. compared users' performance in a find-and-select task using a joystick controlled "static peephole", a dynamic peephole and a handheld AR interface (called Magic Lens) [23]. The study showed that the dynamic peephole and Magic Lens clearly outperform joystick navigation. Morrison et al. compared handheld AR for a paper map with a digital map [18] in a real world setting. One of their conclusion is that the "main potential of (handheld) AR maps lies in their use as a collaborative tool".

A number of studies showed encouraging results for handheld AR systems compared to traditional approaches and researchers investigated fundamental aspects, such as the adaptation of Fitts' law. Much less work addresses the interface design of those systems. Henze et al. proposed different interface designs for a system that augments music records based on a user study [11]. Liao et al. implemented selection and interaction techniques for printed conference proceedings [16]. Based on an initial study they identified challenges for future work (e.g. slow image recognition and inaccurate document registration). What is missing is an understanding of the alternatives of the interface design beyond qualitative results.

3 Interaction with Printed Photo Books

In our work, we investigate solutions that close the gap between printed documents and the digital realms. A particular application area that we currently focus on is the interaction with printed photo books [10]. Photo books are enduring artefacts handcrafted by users to preserve memories and often have online galleries as natural counterparts. To create a connection between the physical and the digital we try to enable

users to access the information available in photo communities' galleries using their physical photo books. By combining digital content, such as comments, music, or videos, with the printed photo book, the photo book could be brought to life.

Photo books are often used by individuals or small groups possibly to tell about the last vacation, the good old times or more generally to share memories. Frohlich et al., for example, found that photos are mainly used to share memories and that "sharing photos in person was described as most common and enjoyable" [8]. People use the photos to "share the memory", and Frohlich also reports that people use printed photos for "jointly 'finding' the memory". Crabtree et al. emphasize that the sharing of printed photos relies upon the distribution of photos across group views and personal views [4]. The following scenario sketches a typical situation that could emerge around photo books:

> *Mary has invited her family to celebrate her birthday. After welcoming all guests and receiving birthday gifts Mary digs out a couple of photo books she produced for recent events such as her trip to Mexico, her son's baptism, and her daughter's school enrolment. A lively discussion about the events emerges between the guests and the photo books are passed from one to the other. While Mary is chatting about Mexico with her brother, her grandparents discuss the baptism in detail and her nephews and nieces debate about the school enrolment.*

While the photo books are a great means to share memories as described above, they come with a few limitations. Content has been selected to create a visual appealing and affordable photo book. More information, e.g. the names of the persons on the photos or comments, as well as additional photos and videos are often available at production time. This background information remains inaccessible using the plain photo book. By augmenting the photos each user could access additional information that are of particular interest to him or her in the specific situation. Mary's brother might want to learn where the photos have been taken exactly and get more information about the photographed sights. He might also be interested in the additional video snippets Mary took during her vacation. Digital content evolves and different users can annotate the content with additional information. Mary's grandparents are, for example, interested in who attended the baptism. Mary would be happy if her parents could provide more information about distant relatives on the photos and if they would add names to the photos printed in the book. Mary's nephews and nieces are probably less interested in the family events but want to get entertained. An audio drama prepared by Mary could connect the photos to an age-specific funny storytelling of the event turning the photo book into an audible picture book for kids.

4 Requirements and Participatory Design

In order to design the handheld AR interface for printed photo books we started with an initial study to collect features and proposals for the interface design. We collected information and features participants want to access using their printed photo book. Furthermore, we asked participants to propose designs to visualize information with handheld AR using pen and paper.

4.1 Methodology and Participants

In the beginning of the study we introduced the participants to the studies purpose and collected demographic data. The remainder of the study was split into two halves. In the first part we asked participants to write down a list of information and features they consider relevant. Participants were also asked to rate each of the named features on a five point Likert scale (from not important to very important). Afterwards, we asked the participants to draw an augmentation on printed images containing a mobile phone that shows an unaugmented image of a photo book on its screen (see Figure 1). The layout of the used photo books is consistent with the sparse knowledge gained from analysing photo books [24] and is also consistent with image composition algorithms for photo books [2].

12 persons (8 male) participated in the study. The participants' age was between 8 and 54 years and the average age was 32.6 years. 4 participants had a technical background (undergraduate and graduate students) and 8 had no technical background.

4.2 Results

In the following we report the results from the user study. First the results for the desired features and information are described, followed by an outline of the sketched interfaces.

Information and Features. The participants named 6.83 (std=3.43) different information/features on average. We normalized the results by merging synonyms and very similar answers. Table 1 shows the average rating for aspects that have been mentioned at least three times.

Table 1. Frequency and rating of named information

Information	*Mentioned*	*Rating*	*Information*	*Mentioned*	*Rating*
persons' names	9	4.2	object description	5	3.4
recording time	8	4.3	comments	4	3.5
recording date	7	4.7	tags/categories	3	4.0
recording place	6	3.8	related images	3	3.5
title/description	6	3.3	links/social networks	3	3.0

The results can be further reduced considering that recording time and date is very similar information that is often presented side-by-side. Persons' names and object descriptions are also similar information that describes specific parts of a photo. Most participants proposed to not only display information but requested the possibility to also create or change additional content. In particular, participants wanted to add descriptions of persons, sights and other objects to photos similar to the way photos can be annotated in online galleries on Flickr and Facebook.

Visualization. The visualizations proposed by the participants can be differentiated by the way the overlay is aligned. Six participants aligned the information to the

border of the phone. Figure 1 (left) shows an example of a sketch where the information is presented at the phone's border. The information is located at either one or two sides of the display. In contrast, three participants aligned all information to the photo. Figure 1 (right) shows one of these sketches that presents the information at the photo's border. The other participants choose a mixed design where some information is aligned to the phone's border and other information is aligned to the photo itself. Particularly, information that describes only parts of a photo is aligned to this part while general information about a photo, such as its title, is located at the border of the phone's display.

Ten participants explicitly suggested highlighting the recognized photos of the photo book in some way. Seven participants proposed to draw a rectangle around the photos. Other participants suggested to gray out the background or did not specify a particular way. Even though not requested, six participants proposed to have a way to activate additional functionalities on a separate view. Two participants proposed using the phone's menu button and two proposed to use icons (e.g. a video icon) that lead to the separated view. The other two participants did not specify a particular way to activate the additional functionalities.

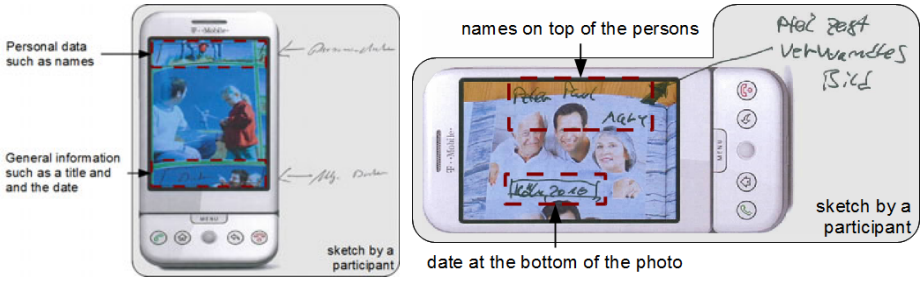

Fig. 1. Sketches of handheld AR interfaces by two participants. Information is either aligned to the display's top and bottom (left) or aligned to the photo (right). For illustrative purpose we selected a very simple sketch.

4.3 Discussion

The information that has been requested most often by the participant describes the photos content such as, persons and objects that have been photographed. When and where a photo has been taken is almost equally important. More general information textual description of a photo such as tags and a title ranks third. Participants could also envisage including social features such as comments or a pointer to social networks. In general, it can be differentiated between information that describe particular parts of a photo (e.g. the name of a person), information that describes the whole photo (e.g. recording time or a title), and content provided by other users (e.g. comments). Most participants do not only want to view information but also want to add information. They propose to be able to select regions of a photo to tag persons, sights, and other objects.

The sketches for the visualization of the augmentation produced by the participants revealed three different pattern that are consistent with a the results of a similar study Henze et al. conducted for augmenting physical music records [11]. Participants align the elements either to the augmented object or they align them to the phone's border. Furthermore, some participants propose mixed designs. Most participants propose to highlight the augmented object. While not all participants proposed the highlighting of augmented objects explicitly in both studies we assume that this a general demand.

5 Design and Implementation

Based on the results of the initial study we designed an interface to interact with printed photo book. In addition to a pure visualization of information we decided to also design interactions to select regions of photos, in order to tag persons, sights, and other objects, as this has been requested by the participants. In the following, we describe the design of the augmentations and the designs of the selection techniques. Afterwards, the implementation of the resulting prototype is outlined.

5.1 Augmentation Design

The proposed designs for the augmentation can be divided in those were the placement of information is aligned to the phone and those were the information is aligned to the object. Furthermore, some participants proposed a mixture of both approaches. We decided to not design a mixed augmentation in order to investigate the alignment aspect without the ambiguity of a mixed design. The participants proposed a number of information items they consider important to visualize. The information can be distinguished in information that describes a whole photo and information that describes a certain part of a photo. In order to support both types we decided to support a title for the photos and annotation of persons and objects in a photo. The title is a representative for information that describes the photo in general while the annotation of persons and objects is the representative for information that describes certain parts.

As proposed by the participants photos and annotated regions are highlighted. For all visualization and annotation techniques the pages of the photo book are highlighted by displaying only the page with colours and leaving the surrounding greyed out. Furthermore, individual photos and annotated regions of a photo are highlighted by drawing a rectangle around them. The centre of the display is marked with a crosshair.

Photo-aligned Augmentation. This design, shown in Figure 2 (left), attaches the information to the photos. The title of a photo is aligned to the top of the photo while the descriptions of particular parts are aligned to the top of a rectangle around this part. The augmentation follows the movement of the photo inside the camera's video. If multiple photos are visible all photos are highlighted and each has its own elements visible simultaneously.

Fig. 2. The photo-aligned interface design aligns the annotations to the photos and for all visible photos simultaneously

The advantage of this design is that all information is visible at the same time. As the text is shown as an augmentation the position and the size of the text changes according to the augmented object. Thus, the size of the presented text can naturally be increased and decreased by changing the distance of the phone to the photo book. One disadvantage of this design is that the augmentation becomes small if the phone is far from the photo. Thereby, text might become difficult to read if a user wants to get an overview about the information available for a photo book page. Furthermore, the text permanently moves and wobbles if the user moves the phone. Thereby, readability of text is affected by accidental movement of the phone.

Phone-aligned Augmentation. The second design, shown in Figure 2 (right), aligns the information to the phone. The photos' titles are at the top of the screen while the

Fig.2. The design shows the annotations aligned to the display's top and bottom. Only the annotations of the photo below the crosshair are visible.

descriptions of particular regions are at the bottom. To connect the textual description of a region with the region in the photo a rectangle is draw around the region and a line connects the description with the region. Thus, the text is always located at the same position. Information is only displayed for the photo that is located below the crosshair that marks the centre of the screen. Thus, even if multiple photos are visible in the camera image only the information for one photo is displayed.

This design has the advantage that the text has always the same readable size. As the text stays at the same position as soon as the crosshair is above a photo readability is not affected by the movement of the phone. The design's disadvantage is that the information for only one photo is visible at a time. Thereby, the user must move the crosshair across all photos on a photo book page to get an overview about the available information. Since the text does not change its position selecting it would be less affected by accidental movement of the phone.

5.2 Interaction Techniques to Select Regions

As annotating regions has been requested in the initial study we designed three selection techniques to mark regions of a photo. As we did not collect recommendation for designing this interaction from participants we designed three fundamentally different approaches. With the first two selection techniques users select regions in the reference system of the augmentation. They either by have to move the phone or touch on the display. We included a third technique where regions are marked by touching a separate static image as a baseline. We did not include the selection techniques Liao et al. proposed with the PACER system [16] because we aim at true handheld AR instead of loose registration and we cannot exploit knowledge about distinct document regions (e.g. words and sentences). The two techniques touch-based techniques can, however, be seen as the basic concepts that are combined in PACER.

Crosshair-Based Region Selection. With the first technique the user aims with the crosshair that is located in the centre of the display at a corner of the region that should be selected. The technique is illustrated in Figure 3. The technique is inspired by handheld AR systems that use a crosshair in the centre of the screen to select predefined objects (e.g. [22]). By touching the display at any position the user defines the first corner (e.g. the top-left corner) of the region. The user than has to move the crosshair to the opposite corner (e.g. the bottom-right corner) by physically moving the phone while touching the display. The region is marked when the user stops touching the screen. As the region is created in the reference system of the augmentation the created rectangle is aligned to the photo.

The advantage of this technique is that the "fat-finger problem" (i.e. that users using touchscreens occlude the area they want to touch) is avoided. As the location at which the crosshair aims can be estimated more precisely than the position in which a touch results it might also be more precisely. As the user can zoom (by changing the distance to the photo) while moving the crosshair, the region could be created more precisely. A disadvantage is that the device must be physically moved. This could lead to a higher physical and mental demand and makes the technique prone to accidental movement of the phone.

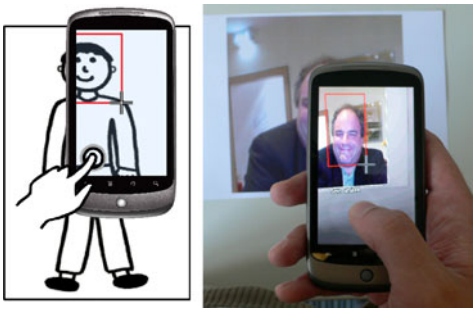

Fig. 3. The three techniques to select regions: crosshair-based on the left, touching in the augmentation in the centre and using touch and a static picture on the right

Augmented Touch-Based Region Selection. With the second technique, shown in Figure 4, the user touches at a corner of the region that should be marked to define the first corner. After moving the finger to the opposite corner the region is marked when the user lifts the finger from the screen. As the region is created in the reference system of the augmentation the created rectangle is aligned to the photo even if the user rotates the phone. This technique is inspired by handheld AR systems where users have to touch the augmentation of predefined objects to select them (e.g. [12]).

Fig. 4. The three techniques to select regions: crosshair-based on the left, touching in the augmentation in the centre and using touch and a static picture on the right

The advantage of this interaction technique is that the user does not have to physically move the phone. Therefore, it might be less physically and mentally demanding. The user can, however, imitate the crosshair-based approach by not moving the finger but only the phone. In this case the techniques have almost the same advantages and disadvantages as the crosshair-based approach. In any case the technique is affected by the "fat-finger problem".

Unaugmented Touch-Based Region Selection. The third technique, shown in Figure 5, serves as a baseline that works on a static image of the photo that should be annotated. As with the previous technique the user touches at a corner of the region that should be marked to define the first corner. The user than has to move the finger to

Fig. 5. The three techniques to select regions: crosshair-based on the left, touching in the augmentation in the centre and using touch and a static picture on the right

the opposite corner. The region is marked when the finger is lifted from the screen. We did not use more sophisticate selection techniques because the other techniques would benefit similarly from more sophisticated interaction techniques.

The clear advantage of this technique is that the user can freely move the phone as the phone is disconnected from the physical photo. Thus, unintentional jitter is avoided but using a separated view can be an imminent disadvantage for a handheld AR system. Another limitation is that the user occludes the area where she aims at with the finger. For this concrete design (but not generally) a further limitation is the lack of zoom.

5.3 Prototypical Implementation

In order to implement the designed visualization and interaction techniques we needed to implement an Augmented Reality system for mobile phones. AR systems estimate the pose of the augmenting display in relation to the scene or object that should be augmented. Using this pose a system can transform the augmenting overlay into the reference system of the physical scene and render the augmentation.

As we aim to preserve the photo book's visual appeal visual markers (e.g. QR-codes) cannot be used. Text-based document recognition [6, 14] is also not possible because photo books contain mostly images. Another widely used techniques to implemented AR systems are recognition algorithms such as SIFT [17]. SIFT and similar approaches are still too demanding for today's mobile phones. Wagner et al. simplified the SIFT algorithm to make the estimation of an object's pose feasible on mobile phones [27]. Their approach is able to process camera frames with a size of 320x240 pixels at a rate up to 20Hz. They further extended the algorithm by combining it with object tracking [27]. This extension enables to recognize and track up to 6 objects with 30Hz. However, only results from processing 6 images are reported and it was not analyzed how the algorithm performs with an increasing number of objects.

In order make handheld Augmented Reality feasible for printed photo books that can contain more than 50 pages we extended the approach by Wagner et al. Similar to [12] we integrated a Vocabulary Tree [19] in the object recognition pipeline. In the pre-processing phase, photo book pages are analyzed to extract simplified SIFT features [27]. Furthermore, the photo's metadata including titles and annotated regions is

converted to a XML format that describes the content of one photo book. Installed on the phone the prototype reads the content description, the according features, and scaled versions of the photos (256x256 pixels). During runtime simplified SIFT features are extracted from the images delivered by the phone's camera and compared to the SIFT features from the photo book pages using the Vocabulary Tree and brute-force matching. If the number of matches is above a certain threshold an according homography is computed. This homography is used to draw an overlay on top of the camera image. To increase the speed, recognized pages are tracked (see [27]) in subsequent camera images. We implemented the algorithm for the Android platform using C (for the performance critical parts) and Java. The prototype recognizes objects in a 320x240 pixel camera frame with 12 FPS and tracks objects in subsequent frames with about 24 FPS on a Google Nexus One. [27] provides an extensive description of their approach and its performance, considering registration errors, and frame rate. We do not use the same implementation but we are certain that the performance is very similar.

Based on the implemented handheld AR algorithm we designed an application that provides the two visualization techniques and the three different ways to select regions. Switching between the visualizations and selection techniques is performed using a menu.

6 User Study

In order to compare the visualization and selection techniques developed in the previous section, we conducted a user study that is described in the following. In the experiment participants performed one task to compare the visualizations and one task to compare the selection techniques. A within-subject design with one independent variable (two conditions in the first task and three conditions in the second task) was used for both tasks.

6.1 Procedure

After welcoming a participant we explained the purpose and the procedure of the study. Furthermore, we asked for their age and noted down the participant's gender. Prior to each task we demonstrate how to use all conditions.

In the first task, participants had to answer five questions related to the photos in a provided photo book. To answer a question they had to read the augmentation shown on a mobile phone. Participants had to combine the information provided by the photos with information provided by the augmentation. E.g. one question was "Who watches soccer?". For this example participants must identify the photo with persons watching soccer and read the annotation that contains the persons' names. After answering a question participants were asked the next question. After completing all questions with one visualization technique they repeated the task with the other visualization and another photo book. We asked participants to answer the questions as fast as possible. The order of the conditions and the order of the used photo book were counterbalanced. We measured the time participants needed to answer the five questions. Furthermore, we asked them to fill the NASA TLX [9] to assess their subjective task load and the "overall reactions to the software" part of the Questionnaire for User Interaction Satisfaction (QUIS) [3] to estimate the perceived satisfaction.

In the second task, we asked the participants to select regions on provided photos. With each of the three selection technique the participants had to mark a region in three photos (e.g. "Mark the person's face."). They could repeat marking a region if they were not satisfied with the result. Participants were asked to mark the region as fast and precisely as possible. After completing the task with one selection technique participants repeated the task with the next technique and a new set of photos. The three conditions were counterbalanced to reduce sequence effects. We measured the time needed to mark each region, the coordinates of the region, and how many attempts participants needed. Furthermore, we asked participants to fill the NASA TLX and the "overall reaction" part of the QUIS.

6.2 Participants and Apparatus

We conducted the user study with 14 participants, 6 female and 8 male, aged 23-55 (M=31.21, SD=8.6). Five subjects had a technical background (mostly undergraduate students) none of them was familiar with handheld AR or the used application.

The prototype described in Section 5.3 running on a Google Nexus One was used for both tasks. The investigator selected the visualization and interaction technique between the tasks. For the first tasks we prepared two photo books printed on A4 and annotated each of the containing photos with a title and/or regions of the photo describing parts of it. The theme of the first photo book was a wedding and the theme of the second photo book was the visit to a fun fair. We prepared an additional photo book for the introduction with photos taken at a scientific conference. For the second task we printed 20 photos on A4.

6.3 Hypothesis

For the first task we predicted that the photo-aligned presentation is more usable than the phone-aligned presentation. With the photo-aligned presentation the user can see all information simultaneously and can quickly focus on different texts by changing the distance of the phone to the photo book. Therefore, we assumed that participants perceive this condition as less demanding and give it a lower NASA TLX score. Due to the same reasons we assumed that participants would give a higher QUIS score to the photo aligned presentation.

For the second task we assumed that the crosshair-based technique would receive a higher QUIS score and that this condition is perceived as less demanding, which would result in a lower NASA TLX score. We assumed that because, compared to the other conditions, the crosshair-based technique can be used with a single hand and the user can zoom and change the selection simultaneously just by moving the phone. For the touch-based techniques we assumed that unaugmented touch would be more usable because the movement of the hand does not move the image that should be selected.

6.4 Results

After conducting the experiment we collected and analyzed the data. We found significant differences between the two visualization techniques as well as between the three selection techniques. We did not find significant effects on the time participants needed to complete the tasks. Participants' qualitative feedback was translated to English.

Augmentation Design. Comparing the two visualization techniques we found that the augmentation design had a significant effect ($p<.05$, $r=0.81$) on the NASA TLX score (see Figure 6). The perceived task load is lower ($M=103.64$) if the augmentation is aligned to the photo compared to the augmentation that is aligned to the phone's border ($M=117.86$). The augmentation design also had a significant effect on the participants average rating of the QUIS's "overall reactions to the software" part ($p<.001$, $r=0.70$). On average the rating is higher if the augmentation is aligned to the object ($M=6.60$) compared to the score for the phone-aligned visualization ($M=5.36$). The individual scores are shown in Figure 6. The visualization technique had a significant effect on the results of all questions ($p<.001$ for the first three questions and $p<.05$ for the others). Task completion time using a photo aligned augmentation is $M=251$s ($SD=95$s) and $M=270$s ($SD=127$s) for the phone aligned augmentation but the difference is not significant ($p=0.07$).

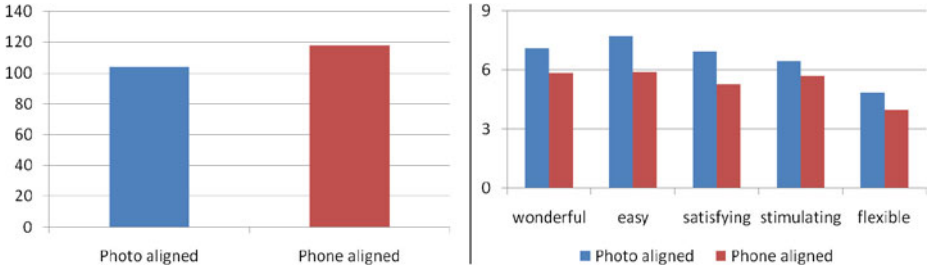

Fig. 6. NASA TLX (left) and QUIS "overall reactions to the software" part (right) for the visualizations

Most of the participants' comments addressed the performance and the accuracy of the object recognition. E.g. one participant mentioned that "it's shaking - probably I hold the camera wrong" and another participant stated that "the recognition should be faster" and the system "should tolerate bended pages". Participants mentioned for both conditions that the recognition works better than with the other condition.

We observed for both conditions that participants prefer to hold the phone sideways. That led to negative comments about the phone-aligned presentation. E.g. "it's difficult to read because the text is skewed" or "have to turn the phone to read the

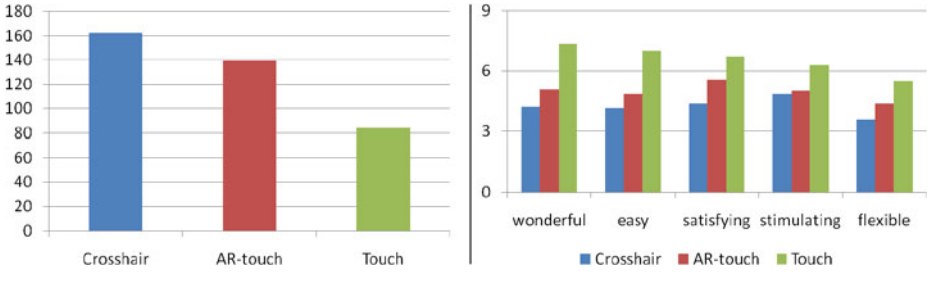

Fig. 7. NASA TLX (left) and QUIS "overall reactions to the software" part (right) for the selection techniques

text". About the photo-aligned condition participants mentioned that "it provides a good overview" and "you can see everything". However, they also mentioned that this presentation is "a bit overloaded" and "you have to go near to use the functionality".

Selection Techniques. In the second task we compared the three selection techniques. We used the Bonferroni correction to correct the significance levels. The analysis of variance (ANOVA) shows that the selection technique had a significant effect on the NASA TLX score ($p<.01$). Comparing the individual condition (see Figure 7) shows that using unaugmented touch ($M=84.07$) results in a lower score than using augmented touch ($M=139.36$, $p<.01$) or the crosshair-based technique ($M=161.79$, $p<.001$). The score for augmented touch is lower than for the crosshair-based technique but, considering the corrected significance level, the effect is not significant ($p=0.025$).

An ANOVA test shows that the selection technique also had a significant effect on the average QUIS's "overall reactions to the software" part ($p<.001$). Using unaugmented touch leads to a higher score ($M=6.57$) compared to augmented touch ($M=4.97$, $p<.01$) or the crosshair-based technique ($M=4.23$, $p<.001$). The score for the augmented touch technique is also higher than the score of the crosshair-based technique (see the individual scores in Figure 7) but without a significant effect ($p=0.027$). Average task completion time for the selection subtasks are crosshair: $M=5.4$s ($SD=4.3$), augmented touch: $M=6.8$s ($SD=5.1$), and touch $M=4.1$s ($SD=2.2$) but the differences are not significant (ANOVA: $p=0.08$).

Even though, we demonstrated the techniques prior the task and ask the participant if he/she understands the technique, some participants did not understand the crosshair-based technique. One participants, for example, noted that "it is difficult to touch the crosshair" although it is not necessary to touch it. Mentioned reasons why this condition performs worse than the others are because it is an "unusual interaction" and that it is "difficult to mark a picture by moving the phone". Another participant noted that "moving the whole body is not comfortable". Further comments are that it is "difficult to catch the crosshair where I want it to be" and the same participants stated that "I always forget paying attention to the crosshair". An advantage participants identified is that "the finger does not occlude the object" and that this technique is "usable with one finger".

For the augmented touch technique four participants appreciated that "it has zoom" (compared to the last condition). Compared to the crosshair-based technique they liked that "one can draw the window with the finger". This condition's most often mentioned limitation is that "the device moves when dragging the box" and that "touching changes the position of the phone" or more generally: "it shakes too much for me".

We got mostly positive comments about the unaugmented touch condition. However, participants identified only one advantage of this technique, even though most participants commented on this advantage. They liked that the "image does not move" and that "the image freezed". They also explicitly stated it is "easy to select because it (the image) does not move". The main limitation the participants identified is that "it has no zoom", that "zooming would be nice" and that it is "less precise than the cursor without zoom". Another problem participants mentioned is that "my finger is to fat" or with other words "there is the fat thumb again".

6.5 Discussion

The results of the first task support our hypothesis that the photo-aligned presentation is more usable than the phone-aligned presentation. Participants perceive the photo-aligned presentation as less demanding and are more satisfied. For the second task the results contradicted our hypothesis. Participants clearly prefer the unaugmented touch technique and the main reason is that image that should be selected does not move.

Based on the sparse comments we assume that the photo-aligned presentation is superior because it provides all information simultaneously and therefore helps to get an overview. The user does not have to select an object to get information about it. This is, however, also the main limitation: The photo-aligned presentation technique does not only allow the user to zoom in and out but it is required to do so. On pages with a high density of annotations the amount of text that is hardly readable can be confusing. We assume the results can be transferred to other tasks with a similar or lower object density. For those tasks the text size could be further increased, which makes it even easier to get an overview. For tasks with a considerably higher object density the text size must be adjusted accordingly to avoid overlapping texts. In this case an object aligned presentation will presumably become less usable because the user has to "zoom" often by moving the phone towards the objects.

The participants clearly preferred to select regions in a static image compared to the two techniques that use AR. This result is surprising because the design of the study favoured the two other conditions. No zoom was available even though a number of well established techniques exist to implement zooming for static images. Furthermore, we did not randomize the order of the tasks and using handheld AR in the first task certainly improved the participants' performance for the two AR based techniques. The qualitative feedback is also quite clear. Participants prefer unaugmented touch because they do not have to deal with the augmentation.

The study has two main limitations. The tasks and the setting are artificial in particular for the first task. For the intended use case it cannot be expected that users will search for particular information. Rather, users usually do not have temporal pressure or want to answer specific questions while browsing through a printed photo book. The second limitation, which applies for both tasks, is the short time participants used the conditions. From this perspective, it is even remarkable that all participants could use the conditions of the first task without any problems. Especially for selecting regions more training would certainly improve the performance with the AR-based techniques. However, it is questionable if training can invert the results. Furthermore, users might not be willing to learn using the crosshair-based technique because it is hard to use at least in the beginning.

7 Conclusions and Future Work

We investigated the design of a handheld AR interface for printed photo books. Based on an initial study we designed two visualizations for augmenting photo books with additional information. These designs and three techniques to select regions of printed photos have been implemented. The subsequent user study shows that aligning annotation to the augmented photo is more usable than aligning annotations to the phone's display. We also showed that users prefer to select regions on static images compared to selection using AR.

We assume that the results are transferable to other application domains. E.g. to interact with printed maps and text-heavy documents. Our results suggest that paper-based handheld AR systems should align text to the augmented objects even if this affects the readability. To select regions of physical objects users should not be forced to use the AR visualization. Simple selection techniques using a static image of the object is clearly preferred by the users.

If the results also apply to handheld AR for 3D objects or large objects in general needs further investigation. As next steps we propose to re-examine selection techniques for handheld AR systems. It should be investigated how more complex approaches (e.g. [16]) perform compared to traditional techniques. Furthermore, potential training effects should be studied for visualizing annotations and for select regions using handheld AR.

References

1. Alessandro, M., Dünser, A., Schmalstieg, D.: Zooming interfaces for augmented reality browsers. In: Proc. MobileHCI (2010)
2. Atkins, C.: Blocked recursive image composition. In: Proc. ACMMM (2008)
3. Chin, J., Diehl, V., Norman, K.: Development of an instrument measuring user satisfaction of the human-computer interface. In: Proc. CHI (1988)
4. Crabtree, A., Rodden, T., Mariani, J.: Collaborating around collections: informing the continued development of photoware. In: Proc. CSCW (2004)
5. Davies, N., Cheverst, K., Dix, A., Hesse, A.: Understanding the role of image recognition in mobile tour guides. In: Proc. MobileHCI (2005)
6. Erol, B., Antúnez, E., Hull, J.: HOTPAPER: multimedia interaction with paper using mobile phones. In: Proc. ACMMM (2008)
7. Fitzmaurice, G.W.: Situated information spaces and spatially aware palmtop computers. Communications of the ACM 36(7) (1993)
8. Frohlich, D., Kuchinsky, A., Pering, C., Don, A., Ariss, S.: Requirements for photoware. In: Proc. CSCW (2002)
9. Hart, S., Staveland, L.: Development of NASA-TLX: Results of empirical and theoretical research. Human mental workload 1 (1988)
10. Henze, N., Boll, S.: Snap and share your photobooks. In: Proc. ACMMM (2008)
11. Henze, N., Boll, S.: Designing a CD augmentation for mobile phones. In: Ext. Abstracts CHI (2010)
12. Henze, N., Boll, S.: Evaluation of an Off-Screen Visualization for Magic Lens and Dynamic Peephole Interfaces. In: Proc. MobileHCI (2010)
13. Henze, N., Schinke, T., Boll, S.: What is That? Object Recognition from Natural Features on a Mobile Phone. In: Proc. MIRW (2009)
14. Hull, J., Erol, B., Graham, J., Ke, Q., Kishi, H., Moraleda, J., Van Olst, D.: Paper-Based Augmented Reality. In: Proc. ICAT (2007)
15. International Organization for Standardization: Information Technology. Automatic Identification and Data Capture Techniques - Bar Code Symbology - QR Code. In ISO/IEC 18004 (2000)
16. Liao, C., Liu, Q., Liew, B., Wilcox, L.: Pacer: Fine-grained interactive paper via camera-touch hybrid gestures on a cell phone. In: Proc. CHI (2010)
17. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2) (2004)

18. Morrison, A., Oulasvirta, A., Peltonen, P., Lemmela, S., Jacucci, G., Reitmayr, G., Näsänen, J., Juustila, A.: Like bees around the hive: a comparative study of a mobile augmented reality map. In: Proc. CHI (2009)
19. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: Proc. CVPR (2006)
20. Pielot, M., Henze, N., Nickel, C., Menke, C., Samadi, S., Boll, S.: Evaluation of Camera Phone Based Interaction to Access Information Related to Posters. In: Proc. MIRW (2008)
21. Rohs, M., Gfeller, B.: Using camera-equipped mobile phones for interacting with real-world objects. In: Proc. PERVASIVE (2004)
22. Rohs, M., Oulasvirta, A.: Target acquisition with camera phones when used as magic lenses. In: Proc. CHI (2008)
23. Rohs, M., Schöning, J., Raubal, M., Essl, G., Krüger, A.: Map navigation with mobile devices: virtual versus physical movement with and without visual context. In: Proc. ICMI (2007)
24. Sandhaus, P., Boll, S.: From usage to annotation: analysis of personal photo albums for semantic photo understanding. In: Proc. WSM (2009)
25. Sellen, A., Harper, R.: The myth of the paperless office. The MIT Press, Cambridge (2003)
26. Wagner, D., Schmalstieg, D.: History and Future of Tracking for Mobile Phone Augmented Reality. In: Proc. ISUVR (2009)
27. Wagner, D., Schmalstieg, D., Bischof, H.: Multiple target detection and tracking with guaranteed framerates on mobile phones. In: Proc. ISMAR (2009)
28. Want, R., Fishkin, K.P., Gujar, A., Harrison, B.L.: Bridging physical and virtual worlds with electronic tags. In: Proc. CHI (1999)