ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LX 19 Number 2, 2012

EMPIRICAL EVALUATION OF AUGMENTED PROTOTYPING EFFECTIVENESS

T. Koubek, D. Procházka

Received: November 30, 2011

Abstract

KOUBEK, T., PROCHÁZKA, D.: *Empirical evaluation of augmented prototyping effectiveness.* Acta univ. agric. et silvic. Mendel. Brun., 2012, LX, No. 2, pp. 143–150

Augmented reality is a scientific field well known for more than twenty years. Although there is a huge number of projects that present promising results, the real usage of augmented reality applications for fulfilling common tasks is almost negligible. We believe that one of the principal reasons is insufficient usability of these applications. The situation is analogous to the desktop, mobile or cloud application development or even to the web pages design. The first phase of a technology adoption is the exploration of its potential. As soon as the technical problems are overcome and the technology is widely accepted, the usability is a principal issue. The usability is utmost important also from the business point of view. The cost of augmented reality implementation into the production process is substantial, therefore, the usability that is directly responsible for the implemented solution effectiveness must be appropriately tested. Consequently, the benefit of the implemented solution can be measured.

This article briefly outlines common techniques used for usability evaluation. Discussed techniques were designed especially for evaluation of desktop applications, mobile solutions and web pages. In spite of this drawback, their application on augmented reality products is usually possible. Further, a review of existing augmented reality project evaluations is presented.

Based on this review, a usability evaluation method for our augmented prototyping application is proposed. This method must overcome the fact that the design is a creative process. Therefore, it is not possible to take into account common criteria such as time consumption.

augmented reality, usability, testing, tangible user interface

It is possible to find a substantial amount of augmented reality (AR) applications on different application markets, especially on the Apple App Store and Android Market. Although they have a significant amount downloads, their rating is generally average, and their real usage is even worse. Therefore, it is necessary to discuss why are these applications usually not well accepted in spite of the promising technology used. We are convinced that one of the principal reasons is the application design. Our hypothesis is that design of these applications was focused especially on the augmented reality adoption, not on the primary process (e.g. effective location of a point-of-interest). In this article, we summarized techniques used for testing of augmented reality applications and outlined a general approach that could be used for such testing.

Generally, it is possible to test three main application properties. The first one is whether the application has required features (e.g. searching within content, export). The second one is how are these features implemented (e.g. slow, imprecise). The last property is the usability (is the user able to find required feature easily, are there any unnecessary repeated processes, etc.). These properties are interconnected – e.g. usability goes hand-in-hand with proper implementation of required features. First two properties could be measured by a number of empirical techniques whereas the usability is problematic.

Obviously, there is no completely universal solution. Such method does not exist even within

common applications with graphical user interfaces, therefore, the situation within various tangible user interfaces is even more complex (tangible user interface is a general term used for different user interfaces used in augmented and virtual reality applications). However, also in this area we could design a general approach that could be adapted on a specific project. Following section outlines methods used for usability testing within the AR applications. Further, a testing approach is designed and applied on several case studies – augmented prototyping application and mainstream mobile AR application. Finally, an evaluation of the proposed testing method is presented.

METHODS AND RESOURCES

Dünser *et al.* (2008) made a survey of evaluation techniques used in augmented reality applications and divided techniques into five groups, depending on the approaches and methods of evaluation. Several examples from each group will be presented. Some testing methods could look rather similar (both based on statistical analysis or fulfilling specific tasks), nevertheless, there are frequently significant differences in the way how these techniques are applied on the problem. We strongly recommend studying further referenced articles for details.

Objective measurements

Approaches in this group usually consist of task completion time measurements and accuracy/error rate evaluations. Other examples are focused on scores, movement, number of actions, etc. In general these methods use also a statistical analysis of measured variables, like ANOVA etc. (e.g. Dünser, 2008). Usage depends on the target application.

Kiyokawa (1999) measures user time that is necessary for finding an object in a space. Users are working in a completely virtual or an augmented environment. Wang – MacKenzie (2000) outlines a test where user is not able to see its hands, just an image of a cube that must be moved on an appropriate place. In this case is measured a time required for identification of the task and time of task competition. Even more, the accuracy of placement and the appropriate angle are also measured.

Similar experiment is described in Swan *et al.* (2007) is focused on depth perception evaluation in head mounted display (HMD). Users guessed distance to the virtual object in a scene. Evaluated was the difference between given and real distance in two different scenes. One scene consists of a "blind test". The scene with object was to the user presented in a HMD. Consequently, the user must reach the object blindly.

Subjective measurements

This group is composed of methods that operate with users. These methods are based on questionnaires, subjective user ratings, or judgements. Some of these studies also employ statistical analysis of the results, others are only based on a descriptive analysis (Dünser, 2008).

Subjective measurement used also Xin – Sharlin (2006) during the evaluation of interaction between human and robot. Users were engaged into a game on wolves and sheeps. Game have two modes – in the first one is user controlling the game by himself, in the other mode uses an artificial intelligence for decision making. Group of users was composed of students and academicians. They were shortly briefed about the rules, and then they played the game. Finally, they filled a questionnaire, and they were debriefed. Xin – Sharlin emphasize that, despite the statistical analysis, the results could be influenced by the fact that some user lost and some wons.

In Juan *et al.* (2005) is outlined a project where AR is used for phobia treatment (particularly arachnophobia). Before the session, a diagnostic interview was done to evaluate the patient's scale of fear and the expectations. Further, the patient was exposed to the fear trigger. Finally, a second interview was processed to evaluate the treatment feedback based on subjective feelings.

Evaluation of impression was used also in Cheok *et al.* (2002) where approx. 20 years old users played an AR RPG game (search for a prices, fight with dragon an a witch). On the end of the game they must answer questions about general impression from game, immersion, comparison of tangible user interface with common GUI, especially from the cooperation point-of-view.

Qualitative analysis

These methods include formal user observations, formal interviews, or classification or coding of user behaviour. According to Woods (2006), qualitative researchers are interested in life as it is lived in real situations. Researcher can use hypothetic-deductive methods, where model of what might be is constructed, then test that model in the situation.

This approach was chosen in Hornecker – Dünser (2007). Children reactions on AR book were evaluated. Selected children were from 6 to 7 years old. The book was a pop-up book with markers that allowed to interact with the story. Scene was recorded from the top by a camera and on the story itself was presented on a display. Children worked with the book individually or in pairs. Especially general reactions and nonverbal actions were evaluated.

Further application tested by quantitative analysis is the telepointer that allows the expert operator to work in a remote environment using HMD. Bauer *et al.* (1999) evaluated work video record, questionnaires and notes from the work. Test was focused on cooperation using the telepointer. Operator depended on the image from HMD that was given by a user, and the operator instructed this user.

Usability evaluation techniques

Following examples presents just a small fraction of many possible usability evaluation techniques. Hix *et al.* (2004) describes so called formative usability evaluation based on fulfilling different scenarios. Evaluation is then based on identification of mistakes during task fulfilment. This approach is used also in Voida *et al.* (2005) for test of manipulation with 2D objects in projector/camera based SAR (Spatial Augmented Reality system). Users had several gestures for object manipulation. The test itself was divided into four phases – exploration, gesture testing, teaching the gestures other user and simple tasks (e.g. moving objects). The goal was to identify which gestures will be easily adopted and frequently used. Other example of this evaluation method can be found in Livingstone *et al.* (2003).

Further group of methods is Summative Usability Evaluation. It is based on statistical comparison of several possible solutions. Criteria for "better" application should be clearly stated before testing. Such testing is done when product is almost finished (see Hix *et al.*, 2004).

Moreover, Expert based evaluation method (also called heuristic evaluation or usability inspection) is used very often. The basic idea is that for testing are not used common users, but a group of experts. Problems identified by this method are used to derive recommendations for improving the design of application. This method should be used to identify critical usability problems early in the development cycle.

An example of such approach is an evaluation of access to the archive of wanted persons described in Brienen – Rodseth (2006). Developers formulated two hypotheses and converted them in questions for experts: "Is AR application more suitable for access to the archive than a common application?" and "Is implemented tangible user interface natural?" (user should not focus on the interface). Further, a group of experts was assembled – an expert on AR applications, user interface expert and police expert. After a short introduction, experts work with the application and this process is recorded for further evaluation.

Similar approach is used in Terry *et al.* (2007). Application is focused on transfer of paper architectonical documentation on TUI. Again, there are two groups of experts – architects and academicians. Architects were focused especially on a common work with the projects (manipulation with plans, getting the information from the plan, etc.), the second group was questioned mainly about the general usability of the application.

Informal evaluations

The last group of the method is called informal evaluations (Dünser, 2008). This category is composed of informal user evaluations such as informal user observations or informal collection of user feedback. As is obvious from the description, there is no strict method for such evaluation. Informal evaluations could possibly identify problems that were not taken into account by other approaches; however, it should be used mostly for additional evaluation.

All outlined methods depend on the specific AR application. It is not possible to identify a general method suitable for all solutions. For evaluation purpose is proper to combine several approaches, because there are many factors that could have an influence on the results: amount of expertise, age, gender, left/right handiness, even cultural aspects (in Islamic countries are users used to read text from right to left). Basis for definition of the evaluation technique can be guideline given by Gabbard (2001).

Case study 1: Desktop application for augmented prototyping

The first case study was focused on AuRel application evaluation (described in Šťastný *et al.*, 2011 and Procházka *et al.*, 2011). This application provides a real-time camera stream on an object with markers. These markers are within the application replace by virtual objects. This allows to adjust the model easily during the design process.

Methodology took into account both empirical and qualitative research. Group of objective and subjective measurements (part of the empirical research) is composed of application response time, time necessary to accomplish a simple task, precision, robustness of the application and stability. Time was measured by stopwatch, tape measure was used for distances and angle was measured using the iOS application Angle Meter on the iPhone 4.

Qualitative research was based on evaluation of a group of users fulfilling given tasks using the AuRel application. A technician observed the users during the test. Further, the users were interviewed after the test for approx. 15 min. The content of the interview was especially their opinion on the application user interface, general user friendliness and whether they would like to use it again for a similar task.

Empirical evaluation: feature testing

It is possible to say that the empirical evaluation describes how well is the application implemented from the functionality point-of-view. An experienced technician does the evaluation. Measurements were done in well-lighted laboratory with Logitech Sphere camera and 14 cm square markers. Evaluation was focused on five key criteria. Measurements were done in different distances from the camera and under different angles (see Tab. I and II). Measured criteria were:

1. The response delay – time between movement of the marker and virtual object movement. Times under one second could not be measured and from the user point-of-view insignificant.

I: *Marker plane and camera image axis are under right angle (ideal position)*

Distance	20 cm	70 cm	170 cm	270 cm	370 cm	500 cm
Movement delay	< 1 s	< 1 s	< 1 s	< 1 s	< 1 s	N/A
Marker movement accuracy	1 pixel	1 pixel	1 pixel	1 pixel	1 pixel	N/A
Marker identification	virtual object blinks	OK	OK	OK	virtual object blinks	loss of marker
Registration stability	permanent jitter of virtual object	no jitter	visible jitter for 5 s from 60s period	visible jitter for 30 s from 60s period	permanent jitter of virtual object	N/A
Pose estimation	OK	OK	OK	OK	OK	N/A

II: *Dependency of virtual object jitter and the marker plane angle*

Distance \ angle	90 degrees	45 degrees	30 degrees
70 cm	no jitter	no jitter	visible jitter, duration 50 s in 60s period
170 cm	visible jitter, duration 5 s in 60s period	visible jitter, duration 30 s in 60s period	permanent jitter
270 cm	visible jitter, duration 30 s in 60s period	permanent jitter	permanent jitter

1: *Workstation with input devices (markers for AuRel TUI and keyboard with mouse for Rhinoceros GUI)*

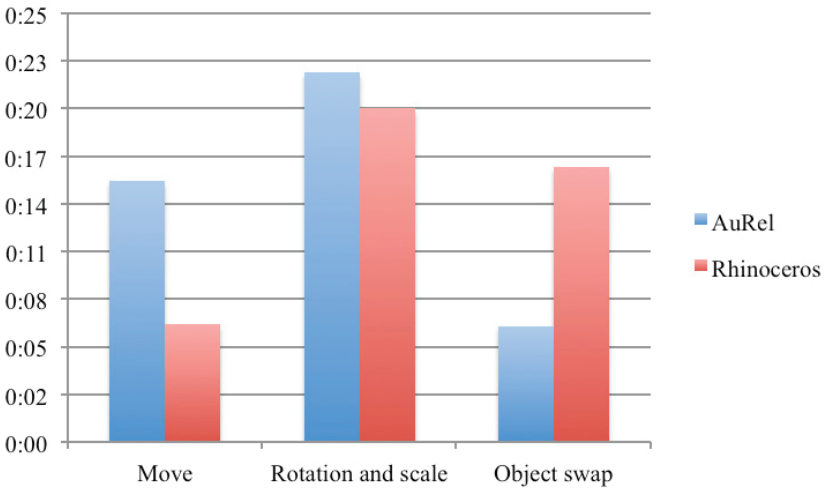
2. Precision of the virtual object placement – correspondence between marker and virtual object position.
3. Marker identification – whether the marker is recognized by the application.
4. Stability of registration – jitter and other disturbing features of the virtual object.
5. Pose estimation precision – whether the object appropriately rotated according to the marker.

In Tab. I is the evaluation of marker detection in different distances from the camera. Marker is always under right angle to the camera axis. Problems with extremely short and long distances are clearly visible. Optimal distance for marker identification is approx. from 50 cm to 200 cm. Precision of the marker alignment is 1 pixel because implemented method is able to identify the object precisely or not at all (other method – SURF – has in this case lesser precision). Therefore, the precision depends on the camera resolution. Tab. II illustrates disturbing effect of lower angles on the marker identification.

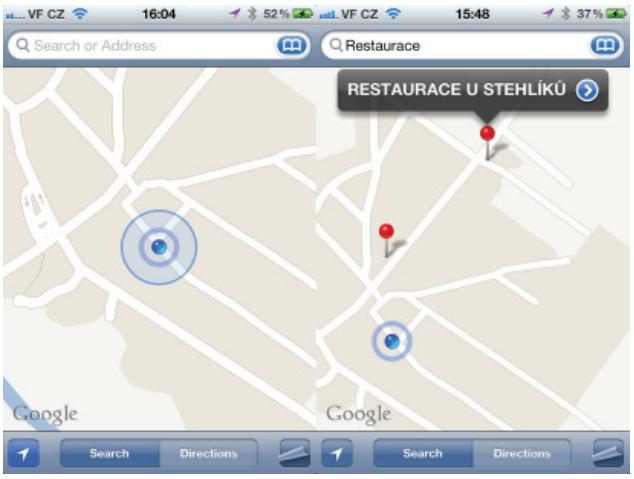
Qualitative evaluation: user friendliness

User friendliness was tested by a group of users composed of students and academicians from different departments. There were 7 men and 2 women. Content of the test was the comparison of AuRel application with Rhinoceros 3D modelling tool. AuRel has just a limited functionality; therefore, we tested just features available in both environments. Rhinoceros has common GUI, therefore, users controlled the application with keyboard and mouse (see Fig. 1). AuRel control was based on described TUI with markers. Both applications were shortly introduced. Further, three common tasks were given – move an object on appropriate position, rotate and enlarge object and switch presented models.

Results are shown in the chart in Fig. 2. We anticipated that the responses time for TUI would be substantially worse in compassion to the common GUI. All user are familiar with desktop applications; however, as been mentioned before, most of them have no experience with AR. Nevertheless, even the second task times are almost same, and the third task is in favour of the AR solution.



2: *Duration of virtual objects tasks (seconds)*



3: *Search for nearby restaurants in Google Maps application on iPhone*

The most significant problems were with the registration of extremely close objects (described by 6 users). Three users complained about confusing concept of looking on LCD while working with some real markers, especially the inability to see the marker from the bottom. Subjectively, users feel that the Rhinoceros is more precise (this is just subjective feeling). Users also complained about the markers size. Two users proposed usage of transparent marker or volume marker. As advantage of the AuRel application was mentioned ability to move with the objects naturally by hand. Generally, the application was described as more intuitive than the Rhinoceros. If users were forced to repeat these tasks frequently, two of them want to use AuRel, four want Rhinoceros, and three some other tool (they are not satisfied with none of them).

The goal of testing was not find a methodology, that proves that AuRel is well designed, for the contrary, we wanted to identify flaws in the design. In many points we succeeded. In qualitative part, the users gave to us recommendations, how to improve the application: they want have handlers on the markers – it decreases risk of marker occlusion by hand, markers and virtual objects pairing editor for users, etc. Even more, results of objective and subjective measurements and the qualitative evaluation are in correlation. Therefore, we can formulate few recommendations:

1. We should improve the marker identification in short distances from the camera.
2. We should improve the marker handling.
3. Possibility of another technique for augmented image displaying should be considered (e.g. head mounted display or spatial augmented reality).

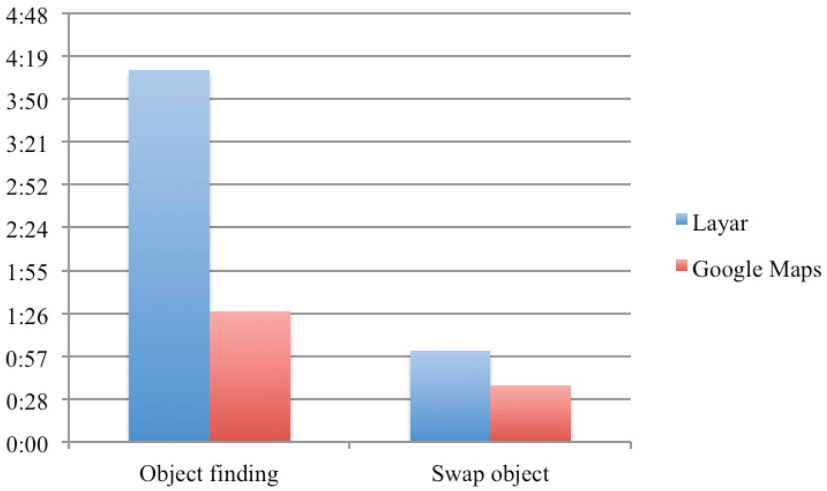
Case study 2: Mobile augmented reality application

In this case study was used the same methodology as in the previous case. The goal is to test whether is not methodology suitable just for a single kind of application. Empirical part was composed of application response time (including the search response time), precision of augmentation, robustness and stability measurements. A stopwatch

4: *Search for nearby restaurants in Layar application on iPhone*

III: *Empirical measurement of the Layar*

Criterion \ Distance	100 m	200 m	400 m	700 m	1500 m
Movement response time	< 1 sec	1 sec	1,5 sec	2 sec	2 sec
Object finding response time	Depends on the internet connectivity, usually under 8 s on EDGE, 2 s on Wi-Fi				
Registration deviation (stability)	< 5 degrees	< 5 degrees	approx. 7 degrees	approx. 10 degrees	approx. 20 degrees
Correctness of registration	Permanent jitter of virtual object				



5: *Duration of task with mobile applications (seconds)*

provided time measurements, distance was measured by TomTom application on iPhone 4 and angles were given by Angle Meter application for iOS on iPhone 4. Qualitative research was based on fulfilling tasks and approx. 15 minutes long interview with the technician (whether is the user satisfied with the application, whether he/she wants to use it again for a similar task, general opinion on user interface).

Part of the evaluation was again comparison of the common application (Google Maps) and state-of-the-art AR application (Layar). Layar provides information about nearby objects from "first person" view (see Fig. 4). Objects are divided into thematic layers. Both application use geo-location services, especially GPS and compass.

Empirical evaluation

Selected results are outlined in Tab. III. Obviously, the Layar application is suitable mostly for navigation to a near object. With larger distance to the destination, grows also the response time and occlusion precision is substantially decreased. Good results are provided up to 200 m distance, acceptable up to approx. 400 m. In further distances, the response delay and imprecision makes the application virtually unusable.

Qualitative evaluation: user friendliness

For this case study we used a group of six people. Three of them were academics, the rest were various family members. They were given two simple tasks: First, to find a specific restaurant in the neighbourhood a clearly state the orientation where the restaurant, then to change the destination to other object. Both objects were from 500 to 1 500 m from the testing position.

Search for a restaurant in Layar application was a substantial problem for all users. None of them was able to identify appropriate layer. After a description of the layer system was one user able to find appropriate layer and required object. To the rest of the users was after 3 minutes selected necessary layer. Time is outlined in char in Fig. 5.

Interviews clearly shown that all users would rather use common navigation such as Google Maps than the Layar. Key problem of the Layar application is the layer system that is strange to the beginners. Very confusing was also the response delay (position of virtual objects was not in correlation with real objects). As a drawback was also mentioned the inability to show navigation information instead of simple orientation. Based on outlined tests, we could formulate following recommendations for Layar developers:

1. Simplify manipulation with layers or remove them at all.
2. Reduce the jitter of virtual object movement in middle and long distances.
3. Decrease the response time during movement with the mobile gear.

DISCUSSION

Our methodology for evaluation of augmented reality application is composed of 2 parts – empirical evaluation and qualitative evaluation. In empirical evaluation we do objective and subjective measurement of criteria from 4 groups – time measurements, accuracy, robustness and stability. Time measurements are responses, delays etc. Accuracy group is composed of measurements of application and user interaction precision. The stability simply means that the application behaves accordingly to the described functionality under all circumstances. Robustness means that a repeated action in the application leads to the same results (in case of same conditions). These criteria must be appropriately adjusted with regard to the kind of application.

Second part of our approach is the qualitative evaluation. In qualitative tests, we focused on user experiences with application. Important is the psychological effect of using application (user can be pleased, surprised, depressed, etc.), user point-of-view on application user interface intuitiveness or friendliness. From results of both case studies we realised, that users could provide very insightful ideas because they are concentrated on fulfilling the tasks, not on some specific technology (as are frequently the developers). Also, the users opinions were mostly in correlation with empirically measured problems.

Therefore, we strongly recommend applying of outlined testing methodology on all projects (not necessarily just on AR applications). The key issue will be to select an appropriate test group. Significant will be especially appropriate age, level of IT knowledge and professional orientation (design application should be tested on designers). Substantial differences could be also in case of repeated tests on the same group. In some cases (special applications) could be learning time longer in favour higher effectiveness later, however, in most cases is necessary to satisfy user immediately (most mainstream applications).

SUMMARY

The article presents a review of different usability testing methods for augmented reality applications. Both techniques based on empirical measurements and so called qualitative evaluation techniques focused mostly on psychological evaluation are described. We proposed a methodology for evaluation of our augmented prototyping application AuRel. This methodology combines some well-known general approaches: objective measurements, subjective measurements, qualitative analysis and user interface testing. We tested our approach on two case studies: desktop augmented prototyping application and mobile augmented reality application for personal navigation. In both cases, the evaluation clearly identified key problems in the application design. Moreover, the qualitative evaluation based especially on interviews with the users was in most cases in correlation with problems identified by the empirical tests. Therefore, we could recommend application of proposed evaluation methodology on similar projects.

Acknowledgement

This paper is written as a part of a solution of the project IGA FBE MENDELU 31/2011 and FBE MENDELU research plan MSM 6215648904.

REFERENCES

KIYOKAWA, K., TAKEMURA, H., YOKOYA, N., 1999: *A collaboration support technique by integrating a shared virtual reality and a shared augmented reality*. IEEE SMC '99 Conference Proceedings.

WANG, Y., MACKENZIE, C., L., 2000: The role of contextual haptic and visual constraints on object manipulation in virtual environments. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* The Hague, The Netherlands: ACM.

GABBARD, J. L.,2001: Researching Usability Design and Evaluation Guidelines for Augmented Reality (AR) Systems. VirginiaTech, Systems Research Center. On-line: http://www.sv.vt.edu/classes/ESM4714/Student_Proj/class00/gabbard/results.html.

XIN, M., E. SHARLIN, E., 2006: Sheep and wolves: test bed for human-robot interaction. In: *CHI '06 extended abstracts on Human factors in computing systems.* Montreal, Quebec, Canada: ACM.

JUAN, M. C., ALCANIZ, M., MONSERRAT, C., BOTELLA, C., BANOS, R. M., GUERRERO, B., 2005: Using augmented reality to treat phobias. *Computer Graphics and Applications, IEEE,* Vol. 25, pp. 31–37.

SWAN, J. E., JONES, A., KOLSTAD, E., LIVINGSTON, M. A., SMALLMAN, H. S., 2007: Egocentric Depth Judgments in Optical, See-Through Augmented Reality. *Visualization and Computer Graphics, IEEE Transactions,* Vol. 13, pp. 429–442.

CHEOK, A. D., YANG, X., YING, Z. Z., BILLINGHURST, M., KATO, H., 2002: Touch-Space: Mixed Reality Game Space Based on Ubiquitous, Tangible, and Social Computing. *Personal Ubiquitous Comput.,* Vol. 6, pp. 430–442.

WOODS, P., 2006: Qualitative Research - Educational Research in Action. *Faculty of Education, University of Plymouth.* [cit. 2011-11-20]. Cited from http://www.edu.plymouth.ac.uk/resined/qualitative%20methods%202/qualrshm.htm.

DÜNSER, A., HORNECKER, E., 2007: Lessons from an AR book study. In: *Proceedings of the 1st international conference on Tangible and embedded interaction.* Baton Rouge, Louisiana: ACM.

BAUER, M., KORTUEM, G., SEGALL, Z.,1999: *"Where are you pointing at?" A study of remote collaboration in a wearable videoconference system*. Presented at Wearable Computers. The Third International Symposium.

HIX, D., GABBARD, J. L., SWAN, J. E., LIVINGSTON, M. A., HOLLERER, T. H., JULIER, S. J., BAILLOT, Y., BROWN, D., 2004: A cost-effective usability evaluation progression for novel interactive systems. In: *Proceedings of the 37th Annual Hawaii International Conference*.

VOIDA, S., PODLASECK, M., KJELDSEN, R., PINHANEZ, C., 2005: A study on the manipulation of 2D objects in a projector/camera-based augmented reality environment. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* Portland, Oregon, USA: ACM.

LIVINGSTON, J. E., SWAN, M. A., HIX, D., GABBARD, J. L., HOLLERER, T. H., JULIER, S. J., BAILLOT, Y., BROWN, D., 2003: Resolving Multiple Occluded Layers in Augmented Reality. In: *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality:* IEEE Computer Society.

BREIEN, F. S., RODSETH, I., 2006: Usability factors of 3D criminal archive in an augmented reality environment. In: *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles.* Oslo, Norway: ACM.

TERRY, M., CHEUNG, J., LEE, J., PARK, T., WILLIAMS, N., 2007: Jump: a system for interactive, tangible queries of paper. In: *Proceedings of Graphics Interface 2007.* Montreal, Canada: ACM.

ŠŤASTNÝ, J., PROCHÁZKA, D., KOUBEK, T., LANDA, J., 2011: *Augmented reality usage for prototyping speed up*. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, LIX, 2, 353–360.

PROCHÁZKA, D., ŠTENCL, M., POPELKA, O., ŠŤASTNÝ, J., 2011: Mobile Augmented Reality Applications. In: *Mendel 2011, 17th International Conference on Soft Computing*. 1. vyd. Brno: Brno University of Technology, s. 469–476. ISBN 978-80-214-4302-0.

Address

Ing. Tomáš Koubek, Ing. David Procházka Ph.D., Ústav informatiky, Mendelova univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika, e-mail: tomas.koubek@mendelu.cz, david.prochazka@mendelu.cz