User-experience in wearable displays: a proposal for standards definition.

The test case of immersive experiencing for user engagement in story telling applications.

Sara Mautino
freelance researcher
sara@lostinreality.net

Mark Melnykowycz
Idezo/Lost in Reality
mark@lostinreality.net

Tags:
augmented reality, user experience, wearable displays, standards, head mounted displays, story telling, geo-location

INTRODUCTION:

The proliferation of internet, computing, and mobile technologies has grown in momentum over the past decade. Communication channels in the internet domain grew from static web pages to dynamic ones, blogging became a web form unto itself, and instant messaging grew from stand along clients to be an integrated as instant updates into social media platforms, which has started to replace email as a standard form of daily communication over the internet. In the past five years, the dramatic rise of mobile computing devices including smart phones and tablets with constant connectivity has made video delivery as second nature as convenient as answering a telephone call. Wearable computing devices, often realized as peripheral add-ons to smart phones are also gaining traction, such as the Pebble Watch, which was a logical evolution following the wide acceptance of small media consumption devices such as the iPod Nano and other micro-sized music players. With the explosion of software and product development strategies across media devices and channels competitors could easily copy and reuse code to build new products, but the differentiating factor between a good and poor product has been defined by a great or poor user experience and the ecosystem of hardware and software.

With the rise of new tools and techniques for creating and distributing media such as text, audio, image, and video, storytelling techniques and ability have also drastically changed, and will continue to evolve. Game design and production for example, now uses many of the same technologies required for movie creation. Novels, which were once consumed in printed forms, can now be read fast on mobile devices, with the ability to mix in other media types such as video and audio to create a more immersive storytelling experience. The next step will be seamlessly including location as a component in the process using transmedia storytelling principles.

This has already begun with location-based games and services that leverage game mechanics such as FourSquare, Findery, Zombie Run, Gbanga, Google Ingress, and other projects. The continuing goal is to leverage current technologies in a way to create a more immersive experience for the person or user consuming that content and being immersed in a storyline.

Augmented Reality devices in the form of Head Mounted Displays (HMD) such as Google Glasses represent another level of synergy between people, content and technology, where wearable devices will always be on and ready to deliver content to the user as a stand alone device, independent from but constantly connected to other devices. A fundamental yet unique aspect in the synergy of a wearable computer with its ‘wearer’ consists in its nature of device enabling context awareness and immersive feeling capabilities. The user becomes user-wearer, the gap between individual and device shrinks and all user-centered related mechanisms acquire a dramatic importance. Given the large array of HMD devices that are being developed, there needs to be a clear way to assess the usability of different devices for specific applications.

Since the pioneering research work dating back to the late 60’s [1], wearable computing developed as a technology driven phenomena. After a decade since their introduction as a real tool in the military battlefield, only in these last years near-to-eye displays found their way to civil applications, thanks to a discrete technological maturity that allows market-readiness in hardware and applications. What is still missing, though, is the existence of a solid conceptual base-structure for testing, measuring, comparing, evaluating wearable displays experiencing under a well defined common standardization system, while having a specific application in mind as a target.
Wearable computers functionalities may be thought of as built up of several blocks. Thad Starner, back in the 1999, in his pioneering thesis work on wearable computing at MIT[2], lists those context-awareness “blocks” into perception (as a collection of context data through sensing equipment), interface, context modeling. It is such modeling, described by Starner as including “observations of the user, the environment and the wearable computer itself” which nowadays still needs to be investigated and developed. Answering questions such as

"What object might be viewable form this room?", "What is the user doing?"[2] is still, after almost 15 years since, an open issue.
In modern terms, what is missing is a framework of Augmented Reality User Experience standard protocols defining an overall conceptual and operative infrastructure for testing and defining observables, units, test bed cycles and user cases (and their relative systems, objects, context)
Focusing attention away from the general augmented reality framework to the more specific domain of HMDs, it is evident the lack of research work and its inhomogeneity.

This lack has been evidenced by three meaningful works published in the years 2005-2008 by Gabbard [3], Swan [4] Dunser [5] and coworkers, who performed statistical analysis about the number of publications appeared in the years 1992-2007 focusing on user-centered studies in augmented reality.
Swan analyzed all peer-reviewed papers from ISMAR, ISWC, IEEE VR conference, plus the MIT Press Journal 'Presence'. Of a total of 1104 initially selected articles, only 266, equivalent to 8%, where found to be AR related: of those, only 21, corresponding to 2%, included a formal UX analysis in AR, a very small fraction of total research in this area.
Swan's work is substantially confirmed by Dunser in a later research. With his team he considered a group of most common publisher databases and selected the ACM Digital Library and IEEE Xplore. out of a total of about 30 publishers, as the most relevant for the area. They counted and cyclically filtered the documents published along the years 1992-2007, ending up with a total of 3309 AR related papers over a total of 6071 general ones. In order to further select the documents related AR papers including user-centered evaluations, they elaborated a selection strategy based on the design of ad-hoc queries and keywords.
The final result was a total of 161 filtered documents in AR domain containing some user evaluation: 10% of the initial 6071.
Olsson properly affirms [6] about user expectations and user experience in augmented reality that "most of the user research that exist has focused on evaluating early technical demonstrators in specific contexts especially looking into perception and cognition issues, user task performance, or other usability-related aspects. User experience and acceptance of such AR demonstrators have been often dismissed, especially when considering emotional aspects of the experience"
So, apart from the scarce attention devoted by the scientific community to the user centered research for AR domain, what matters is that the already existent work is still fragmented, unorganized and specifically oriented according to technologies, methods and contexts.
We think a big effort is needed in the direction of rearranging what has been done till now, and then re-use it in order to create upon the existent state of the art more solid bases for the future. In doing this, it should be kept in mind the importance of the emotional aspect in the user experiences, often dismissed in the technology driven development trend of wearable displays.
Emotion and quality of user perception and engagement play in the end a fundamental role in the effectiveness of an AR product and consequently, in its choice.

EVALUATION METHODS

As an attempt to cope with the issues presented, a starting point is to show examples of relevant methods adopted in literature regarding user centered studies. The existing evaluation procedures may then be elaborated upon and enriched in the direction of building a robust user-centered framework.

In the following, two different approaches are proposed from literature: the first one focuses on subjective aspects of user experience, while the second deals with the analysis of a set of objective variables. A third example goes more specifically into the details of UX evaluation methods for AR.

The selected methods do not pretend to be the best ones available in literature, neither to represent exhaustively a state of the art panorama, rather they were selected as a meaningful and reasonable starting point to be illustrated for the systematic quality in the approach and for the central role given to the tools employed, such as the questionnaires or the statistical analysis.

Disregarding the lack of research work for wearable displays in the user-centered domain, the area is by its nature so fragmented, and the number of relevant aspects and parameters to be included so high, that carrying out a proper critical reviewing action of the state of the art represents certainly a challenge.

Subjective indexes

It is worth here to review three common tools often used for subjective experience classification called the subjective indexes[7]. Visual Symptoms Questionnaire (VSQ)[8], Simulator Sickness Questionnaire (SSQ)[9] and task load index (NASA-TLX)[10] are tools usually proposed to individuals when needing to evaluate media interfaces under the point of view of the subjective reaction to a certain experience.
The VSQ is a questionnaire for the purpose of assessing visual symptoms while the SSQ is a questionnaire designed to determine simulator sickness based on three components (nausea, oculomotor and disorientation) and they are based on a four-point scale response. NASA-TLX is instead a method for evaluating workload with regard to tasks, where six weighted subscales (mental demands, physical demands, temporal demands, own performance, effort and frustration) are combined to determine the total workload score.
The three questionnaires matrixes can be found in the Appendix.

Procedure I

Polonen [11], in 2010, proposes a procedure for comparing non-see through head mounted display systems under a subjective experience point of view. The method can be easily generalized to the see-through version by adding experiments regarding additional subjective and objective testing variables concerning context and HCI: object, context and the system (cite and define).
4 different HMD are compared from the user perspective: EMG iTheater BP4L, MicroOptical MyVu MA-0341, Vuzix iWear AV920 model 242 and Zeiss Cinemizer 1488-603 for 106 individuals, by letting the users watch a movie projected directly into the Near To Eye Display (NED) device and gathering information before and after the experience through the VSQ, SSQ and NASA-TLX questionnaires. The task of 'TV-watching' was included as a control status.
A rigorous statistical analysis is subsequently performed by employing specific non-parametric statistical texts.

All the evaluation processes were adopted according to the general scheme illustrated below and lasted around 1.5 hours.

1. Introduction to the experiment
2. Questionnaires (1) SSQ, VSQ, other questions
3. Vision of the film into the HMD device (TASK)
4. Questionnaires (2) SSQ, VSQ, Other questions
5. Device parameters specifications gathered
6. Statistical analysis: Wilcoxon signed ranks test Kruskal-Wallis test, Mann-Whitney U GLM Univariate Kendall's tau-b

Questionnaires (1) has the function of visual screening background questionnaire before near-to-eye device (NED) exposure: SSQ, VSQ questionnaire and other information requests concerning headache, history, gender, age, motion sickness susceptibility, medication, wearer of glasses, in technology, education, work, computer use, previous experience with games and with NED and virtual environments.
As a visual screening concerning individual characteristics, were also asked questions concerning visual acuity (near and far), interpupillary distance, stereo acuity, color vision, phoria, near point of accommodation.
Questionnaires (2): SSQ, VSQ, NASA –TLX and other information requests concerning: effort, frustration, physical demand total workload, pleasantness, headset fit, visual quality, opinion change, NED-related future interest.

Device parameters specifications were collected directly from manufacturer:

1. weight
2. luminance cd/m2
3. contrast ratio
4. focal distance left (m)
5. focal distance right (m)
6. convergence distance (m)
7. FOV horizontal (°)
8. FOV vertical (°)

9. FOV diagonal (°)
10. QVS horizontal
11. QVS vertical
12. Interocular distance (mm)
13. Luminance difference
14. Vertical misalignment (°)

Statistical analysis.

For the statistical analysis a series of non parametric tests was adopted, according to the non Gaussian distribution of the considered sample population; each of them was chosen depending on the nature of the specific evaluation.

- Wilcoxon signed ranks, test for comparing two sample series of related values, like 'before' and 'after' NED exposure
- Kruskal-Wallis test, for comparing more than two not related samples (i.e. like an opinion change)
- Mann-Whitney U, used for paired comparisons, for assessing whether one of two samples series of independent observations tends to have one significantly larger value than the other. It is one of the best-known non-parametric significance tests)
- GLM Univariate
- Kendall's tau-b , a statistic to measure the association between two measured quantities

The analysis was performed according to 5 classes of comparison: Task Pleasantness, Opinion Change, Visual Quality, and Headset Fit; Workload and Sickness; Comparison of the TV and NEDs; Relationships Between Evaluated Variables; Individual Characteristics correlation.

Objective indexes

By definition, objective indexes are objectively quantifiable or measurable quantities, specific for each parameter of a test case and may include the subject (for example its heart rate), the virtual object (for example the image contrast), the context (attributes such as the luminosity) or the whole system.

Procedure II

Kawai and coworkers [7] propose an experiment to evaluate the UX in an outdoor experiment analyzing two monocular see-around HMDs to watch video content while walking. Additionally, both the case of a hand-held media player and without-stimulation were included in the study as control cases. A group of 8 individuals were involved in the study.
The task consists of the act of walking through a large shopping mall, using escalator when moving between floors, while watching video contents.
Aside from a subjective index estimation, here again based on a statistical analysis of SSQ,VSQ and NASA-TLX questionnaires, in this case, an objective index estimation was also performed.

As objective indexes, the heart rate was measured and given the valence of response to psychological and physical loads, while the walking speed was interpreted as an indicator of the task load; both values were acquired in 5 second intervals by means of a wrist-type module. In addition, the environment in front of the subject's face was video-recorded by a small camera mounted to the center of the goggles, saved on pocked-sized video recorder and then quantified by the counts of the subject's sight switches between looking forward and downward.

1. introduction to the experiment
2. visual acuity and stereovision (Randot Stereotest) assessment
3. Questionnaires (1) SSQ, VSQ
4. walking through mall with HMD (TASK) and experimental acquisitions
5. Questionnaires (2) SSQ, VSQ, NASA-TLX
6. Device parameters specifications gathered
7. Statistics. ANOVA analysis for: heart rate, walking speed, fwd/down sight counts

Four design parameters of were collected from the manufacturer: resolution; virtual distance; screen size; weight.

Statistical analysis.

The statistical analysis was based on the ANOVA analysis of variance.

Testing variables classification

Figure 1 depicts a flow chart of basic user experience parameters for HMD devices. A more specific list of variables are displayed in Table 1. This scheme represents a first attempt in outlining a general HMD standard testing protocol, according to the variable 'space' definition. Listed are possible variables that may play a role into the testing process with a given classification and function. In the future, a complete matrix of elements and instruments could be created and made ready to be implemented in a standard procedure that could be applied to a specific user case for testing.

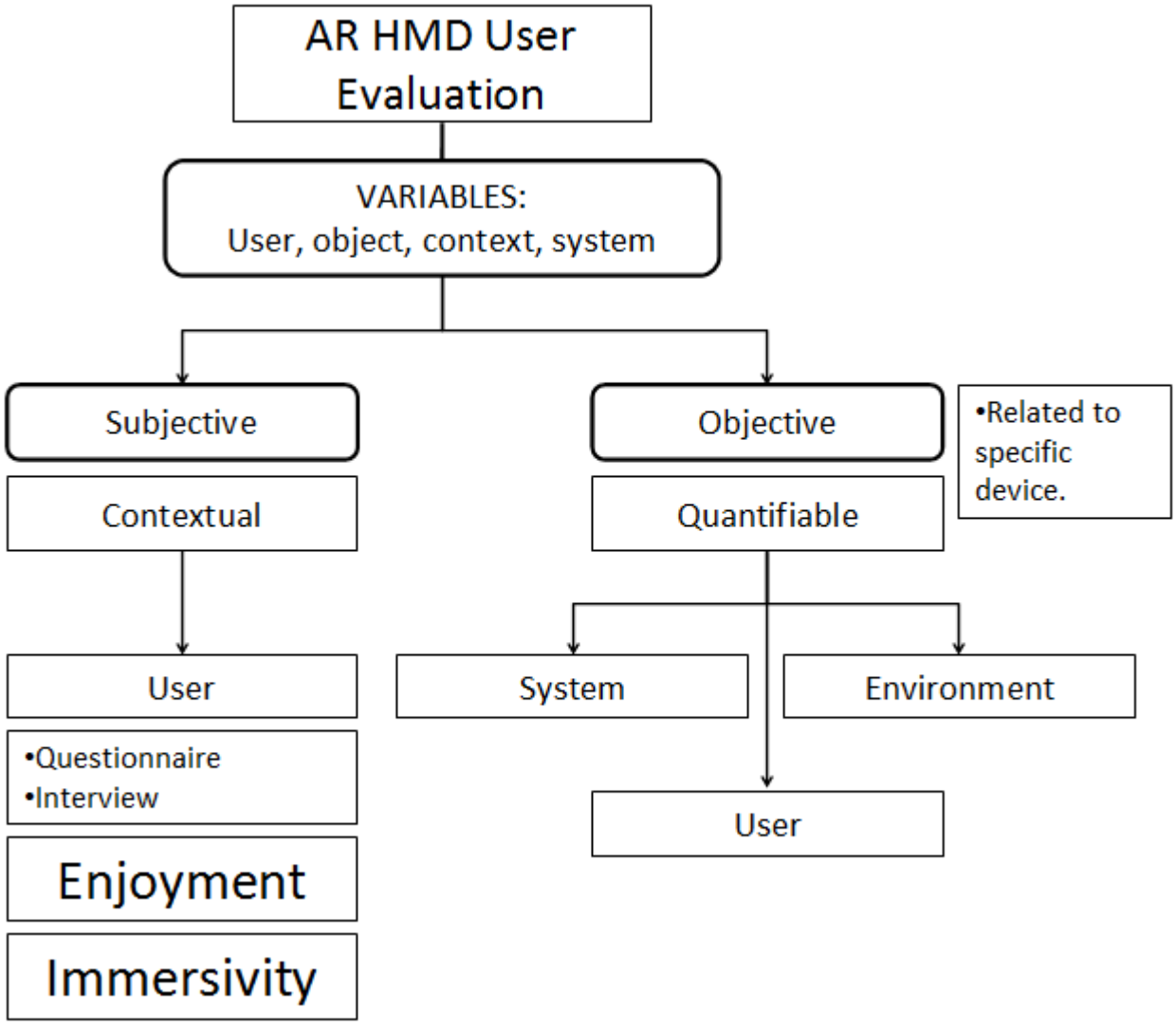

Figure 1 Schematic summary of subjective and objective user evaluation quantities.

Table 1 Listing of user test case and variables from user and device standpoints.

USER TEST CASES / VARIABLES		(users standard test cases [12])
User	Subjective (individual)	User wellness: Diagnosed migraine, Wear glasses, Age, Gender Technology oriented/ interested (early adopter) Game player Minimum Frame-rate sensitivity
	Subjective (task related)	User wellness: nausea, eye strain, headache, max wearable time Ergonomic comfort Test Result
	Objective	User: Heart rate, walking speed, searching, reading, distance judgment, spatial memory, watching video Environment: Luminance vs outdoor light, outside lighting, temperature, noise, distraction elements
Object	Objective	Virtual: Symbology : size, position, visibility, priority, transparency, density Real : size, position, visibility, priority, density (close to context, depending on app)
Context	Objective	User environment, luminosity, ambient noise levels and motion
System: Optics Hardware UI	Objective (specs/ measured parameters)	OPTICS: FOV, occlusion, depth perception Resolution Vignetting Focal distance Luminance Type: Monocular, binocular, bi-ocular, see-through/not see-trough UI: frame-rate, mechanic stability
	Subjective	Aesthetics Comfort
Task		watching video, walking, walking through floor by elevator, responding to information communicated via HMD display
Questionnaire	Subjective	Visual Symptoms Questionnaire (VSQ) Simulator Sickness Questionnaire (SSQ) Nasa Task Load Interviews
Statistical test	Objective	Test such as: Wilcoxon signed ranks test, Kruskal-Wallis test, Mann-Whitney U test, GLM Univariate test, Kendall's tau-b test, variance analysis
Control status device		TV hand-held devices hard copy document no stimulation case

Testing methods classification

Aside from the definition, the system of UX variables must also be structured into a framework of UX testing and classification methods. Bowman, Gabbard and Hix in [13] propose an interesting and quite complete classification of usability evaluation methods. They developed a conceptual UX framework focused on usability issues in Virtual Environments (VE) that could be easily adapted to the wider field of general UX for AR systems and, specifically, to the wearable displays framework. Again we cite a meaningful example, with the aim to use it as a base for the further development of a more complete standard system for User Experience evaluation in HMD.

Evaluation methods are classified according to three main characteristics, and distinguished by the degree of involvement of representative users (user or usability-expert participation), the context of evaluation (generic or application-specific), and the types of results produced (qualitative/quantitative).
Each of the three classes may be seen as the column header of a Usability Evaluation Methods Matrix, where in each box are distributed the existing evaluation tools, such as *Cognitive Walkthrough*, *Formative Evaluation*, *Heuristic or Guidelines-Based Expert Evaluation*, *Post-hoc Questionnaire*, *Interview / Demo*, *Summative* or *Comparative Evaluation*[13]
In the paper it is claimed that, aside from structuring the space of evaluation methods, such a classification enables the estimation of evaluation cost, impact and type of results and provides as well a vocabulary for discussion of methods in the research community.

Another important classification is made according to the type of approach adopted: a distinction is made between the test-bed and the sequential evaluation. The first approach is a general one, and focuses on low-level tasks, while the second is application-specific. A detailed description of the two approaches goes beyond the aim of our work, but the flowcharts of the two are showed below as a summary. A suggested solution is to adopt both approaches in an iterative way.

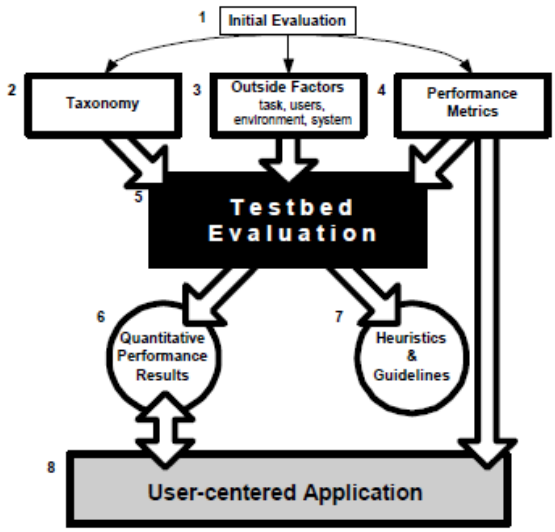

Figure 1a Test-bed evaluation [13]

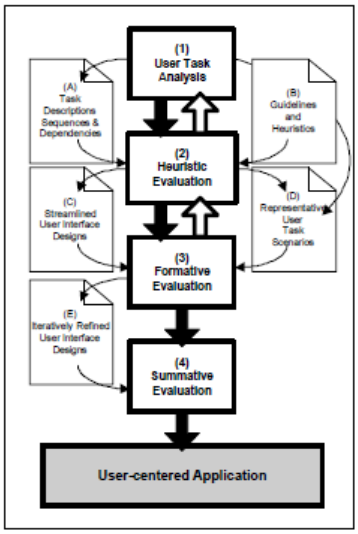

Figure 1b Sequential evaluation[13]

Conclusions

The rise of HMD technology has the potential to transform various user patterns in personal computing. However the technology must be matched to good user experience evaluation and testing standards as new devices and novel applications come on to the market. Historical research into virtual experiences, HMD, and AR devices have not sufficiently addressed the user experience component of device design. A combination of subjective and objective evaluation criteria could be used in the future in the form of an AR UX testing framework to properly evaluate new HMD devices.

One of the aims of the current review of standards and user experience research for HMD devices was to ascertain how the use case of location and mobile storytelling could be implemented for the Lost In Reality

mobile product [14]. Lost In Reality is a mobile application currently in development, which connects story elements (text, audio, video, pictures) to GPS locations, so a user can tell or follow a story through a city. A HMD device may be more ideal than a mobile phone for optimal user experience, presence and immersivity, but to date there is no framework for ascertaining which upcoming devices would be ideal for Lost In Reality.

The comparison has to be performed systematically -i.e.- comparing the user experience by actually 'measuring' and evaluating a set of parameters to be chosen in an appropriate way, performing the evaluation with a rigorous approach: in a controlled scenario, with a set of controlled environment conditions and a well defined testing variables framework.
We think that in the area of wearable displays a fundamental challenge is the elaboration of common guidelines that may help customers, researchers and manufactures to choose and classify HMDs from the user perspective, and to build up a common language - with a proper vocabulary and units system - for information exchange. The necessity is also driven by the huge quantity of new models and prototypes appearing on the scene on an almost daily base and by the lack of systematized research activity carried out in the field of the user experience till now.

To succeed in the task of filling this lack is, in our opinion, a certainly challenging but extremely needed effort.

BIBLIOGRAPHY

[1] I. E. Sutherland, "A head-mounted three dimensional display," in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, San Francisco, California, 1968, pp. 757–764.
[2] T. Starner, "Wearable Computer and Contextual Awareness," MIT, 1999.
[3] "Usability Design and Evaluation Guidelines for Augmented Reality (AR) Systems: Joe Gabbard." [Online]. Available: http://www.sv.vt.edu/classes/ESM4714/Student_Proj/class00/gabbard/index.html. [Accessed: 10-Feb-2013].
[4] J. E. Swan and J. L. Gabbard, "Survey of user-based experimentation in augmented reality," in *Proceedings of 1st International Conference on Virtual Reality*, 2005, pp. 1–9.
[5] A. Dünser, R. Grasset, and M. Billinghurst, "A Survey of Evaluation Techniques Used in Augmented Reality Studies," 2008.
[6] T. Olsson, T. Kärkkäinen, E. Lagerstam, and L. Ventä-Olkkonen, "User evaluation of mobile augmented reality scenarios," *Journal of Ambient Intelligence and Smart Environments*, vol. 4, no. 1, pp. 29–47, Jan. 2012.
[7] T. Kawai, J. Häkkinen, T. Yamazoe, H. Saito, S. Kishi, H. Morikawa, T. Mustonen, J. Kaistinen, and G. Nyman, "Ergonomic evaluation of ubiquitous computing with monocular head-mounted display," pp. 754202–754202, Feb. 2010.
[8] P. A. Howarth † and H. O. Istance, "The association between visual discomfort and the use of visual display units," *Behaviour & Information Technology*, vol. 4, no. 2, pp. 131–149, 1985.
[9] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, 1993.
[10] S. G. Hart, "Nasa-Task Load Index (NASA-TLX); 20 Years Later," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, Oct. 2006.
[11] M. Polonen, T. Jarvenpaa, and J. Hakkinen, "Comparison of Near-to-Eye Displays: Subjective Experience and Comfort," *Journal of Display Technology*, vol. 6, no. 1, pp. 27 –35, Jan. 2010.
[12] PEREY Research & Consulting and AR Standards Community Meeting, "Mobile Augmented Reality Use Cases Position Paper," Mar-2012 [Online]. Available: http://www.perey.com/ARStandards/AR_Guide_Use_Case.pdf [Accessed: 14-Feb-2013].
[13] D. A. Bowman, J. L. Gabbard, and D. Hix, "A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 4, pp. 404–424, Aug. 2002.
[14] "Lost In Reality | Location Based Storytelling." [Online]. Available: http://www.lostinreality.net/. [Accessed: 14-Feb-2013].

APPENDIX - Subjective indexes questionnaires

SIMULATOR SICKNESS QUESTIONNAIRE (SSQ [29])

General discomfort	None	Slight	Moderate	Severe
Fatigue	None	Slight	Moderate	Severe
Headache	None	Slight	Moderate	Severe
Eye strain	None	Slight	Moderate	Severe
Difficulty focusing	None	Slight	Moderate	Severe
Increased salivation	None	Slight	Moderate	Severe
Sweating	None	Slight	Moderate	Severe
Nausea	None	Slight	Moderate	Severe
Difficulty concentrating	None	Slight	Moderate	Severe
"Fullness of the head"	None	Slight	Moderate	Severe
Blurred vision	None	Slight	Moderate	Severe
Dizzy (eyes open)	None	Slight	Moderate	Severe
Dizzy (eyes closed)	None	Slight	Moderate	Severe
Vertigo	None	Slight	Moderate	Severe
Stomach awareness	None	Slight	Moderate	Severe
Burping	None	Slight	Moderate	Severe

Visual Strain Questionnaire (VSQ [23])

Tired eyes	None	Slight	Moderate	Severe
Sore or aching eyes	None	Slight	Moderate	Severe
Irritated eyes	None	Slight	Moderate	Severe
Watering or runny eyes	None	Slight	Moderate	Severe
Dry eyes	None	Slight	Moderate	Severe
Hot or burning eyes	None	Slight	Moderate	Severe
Blurred vision	None	Slight	Moderate	Severe
Double vision	None	Slight	Moderate	Severe
General visual discomfort	None	Slight	Moderate	Severe

NASA-TLX questionnaire

RATING SCALE DEFINITIONS

Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
PERFORMANCE	*Good/Poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?