Usability testing of a Hypervideo-based Interactive TV application

Toni Bibiloni[1], Antoni Oliver[1], Cristina Manresa-Yee[2], Javier Varona[2]

[1] Laboratori de Tecnologies de la Informació Multimèdia. Department de Ciències Matemàtiques i Informàtica. Universitat de les Illes Balears. Spain

[2] Unitat de Gràfics, Visió i Intel.ligència Artificial. Department de Ciències Matemàtiques I Informàtica. Universitat de les Illes Balears. Spain

{toni.bibiloni, antoni.oliver, cristina.manresa, xavi.varona}@uib.es

Keywords: Hypervideo, Interactive TV, Augmented Reality, Second-screen application, Usability testing.

Abstract. In this paper the usability testing of an Augmented Reality system for the Hypervideo Platform, used to simulate augmented reality on Interactive TVs thanks to the hypervideo concept, is presented: a heuristic evaluation was conducted for the creation module, while a within subjects experiment was performed for the visualization module, comparing the previous TV-only system with a second-screen device alternative introduced to improve the usability of the system.

Introduction

A hypervideo, or "video with hyperlinks" [1] is an interactive video stream in which the user is able to interact with the content through hyperlinks, leading to non-linear navigation, searching, sequence skipping, etc. with the purpose of improving the access to the information and bringing the viewer from a passive to an active state [2].

When the hypervideo concept is applied to real images recorded in a video product, augmented reality can be experienced, when this indirect view of the real world is combined with virtual elements, creating a mixed reality.

This paper follows the work from previous papers [3, 4] and demo [5], where a hypervideo platform capable of creating and delivering an AR experience to the viewer through current generation Interactive TV solutions, such as HbbTV, Android TV or Samsung Smart TV was presented. The platform was later improved with a second-screen device option to improve the usability of the visualization module [6].

Whilst a preliminary functionality test was conducted between audiovisual producers and potential viewers (University students) in previous work, no efforts in assessing the usability of the system were done. Now that the functional requirements are clear and have been implemented a double usability test is proposed: a heuristic evaluation for the authoring process, following Jakob Nielsen's heuristics [7], and a within subjects experiment for the visualization module, comparing the previous TV-only system, using the TV remote, with the later introduced second-screen device alternative.

In the first section, the previous work is reviewed to introduce the platform to the reader. Then the usability test is introduced, describing the evaluation processes for the two modules. The paper ends with the conclusions obtained from performing the usability test to the Hypervideo platform.

Previous work

The hypervideo solution

The hypervideo format chosen in this project has three dimensions:

- An audiovisual track, which represents the PoIs and is the base for the whole product. In previous work it was delivered via streaming and now is intended to be played through the broadcast channel.
- The points of interest (PoIs), plus their additional information, which can be textual, visual and complementary.
- The markers that represent these PoIs on the video track that enable the user to identify them (*hot-spot* role) and access its additional information (hyperlink role).

This format was proven to be understood by audiovisual producers in previous work and was used to create functional demos.

The hypervideo platform

The hypervideos are created and viewed thanks to the hypervideo platform. The proposed architecture for the platform is shown in Figure 1, consisting in two modules which interact with a server in the middle.

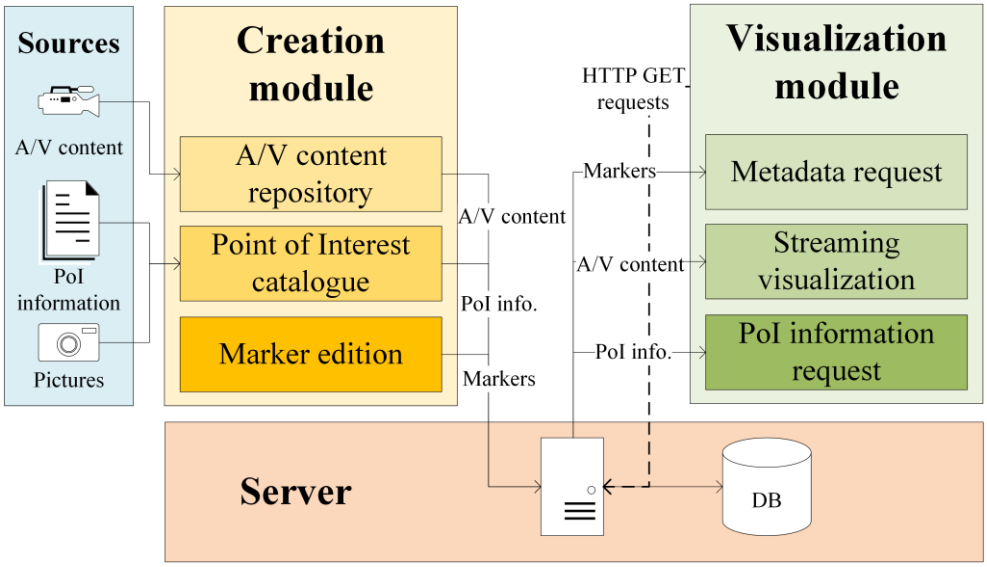


Figure 1. Diagram of the architecture of the hypervideo platform.

The creation module comprises the tools needed to create a hypervideo, starting by managing the audiovisual repository and inserting new data in the PoI catalogue.

Once these steps have been completed, PoIs are linked to the media through the positioning of the markers that represent them with the aid of an interactive tool.

This module has been developed as a Javascript web application hosted in the server.

The visualization module is composed by the hypervideo player applications, with the ability to playback the audiovisual track, represent the markers over it and show the additional information of the PoIs requested by the viewer.

A multiplatform development has been followed, being implemented in HbbTV, Android TV and Samsung Smart TV technologies, using HTML and Javascript.

The hypervideo server is the agent between both modules. It serves the creation module as a web application, and the creation module for the HbbTV and Android TV platforms, as they use a web-based approach.

The server also stores all the data related to the hypervideos and handles the requests from the modules.

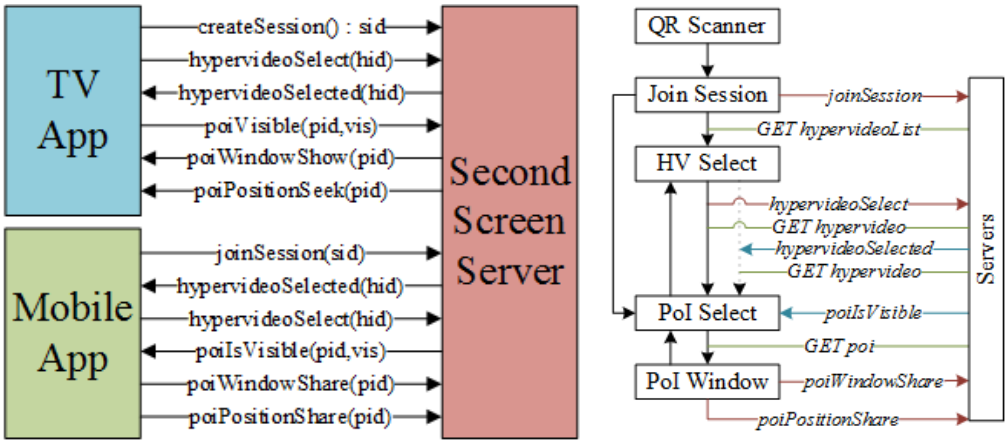


Figure 2. Second-screen application communication protocol & use in the companion app.

A second-screen companion application has been developed for Android devices, which communicates with a second-screen server, which establishes a communication protocol through WebSockets, as described in Figure 2. Both applications, the TV application and the companion app, report events to the second-screen server, which is responsible of forwarding these to the other(s) device(s), as more than one companion app can be used at once.

The user interface

In this section, the user interface of these three systems is reviewed:

- Creation module
- Visualization module: TV application
- Visualization module: companion app

The first one is a web application, which consists of forms to create hypervideos, categories and points of interest, like the shown in Figure 3, and the marker positioning tool, also shown in Figure 3.

A navbar is always displayed at the top, which facilitates navigation through the different forms to create or edit items.

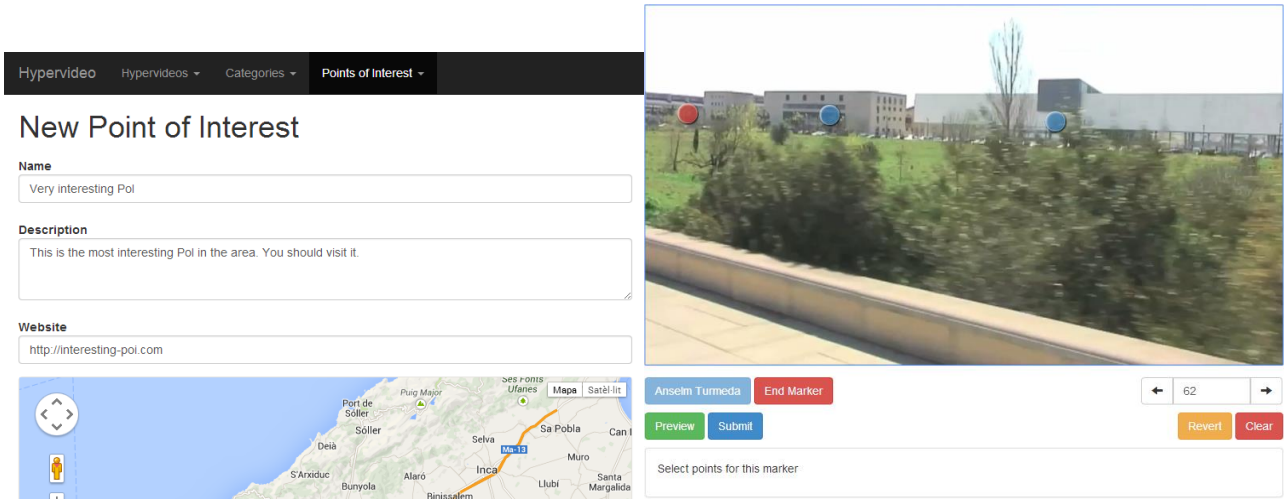


Figure 3. User interface of the creation module: PoI catalogue form, marker placement tool.

The TV application of the visualization module (Figure 4) is meant to be controlled with the TV remote, as it is an Interactive TV application. For this reason, a button reference is always shown in the bottom part of the screen. The first capture shows the hypervideo selection menu, where the user can select a hypervideo to play using the arrows in their remote; then, the hypervideo plays, while the markers display on the top of it, the category selector is on the top right corner; the user may pause and select a marker to obtain

additional information of the PoI, as in the third capture; finally, additional information can be obtained with the arrow keys, such as more pictures, a map showing the location of the PoI or a QR encoding the website of the PoI.

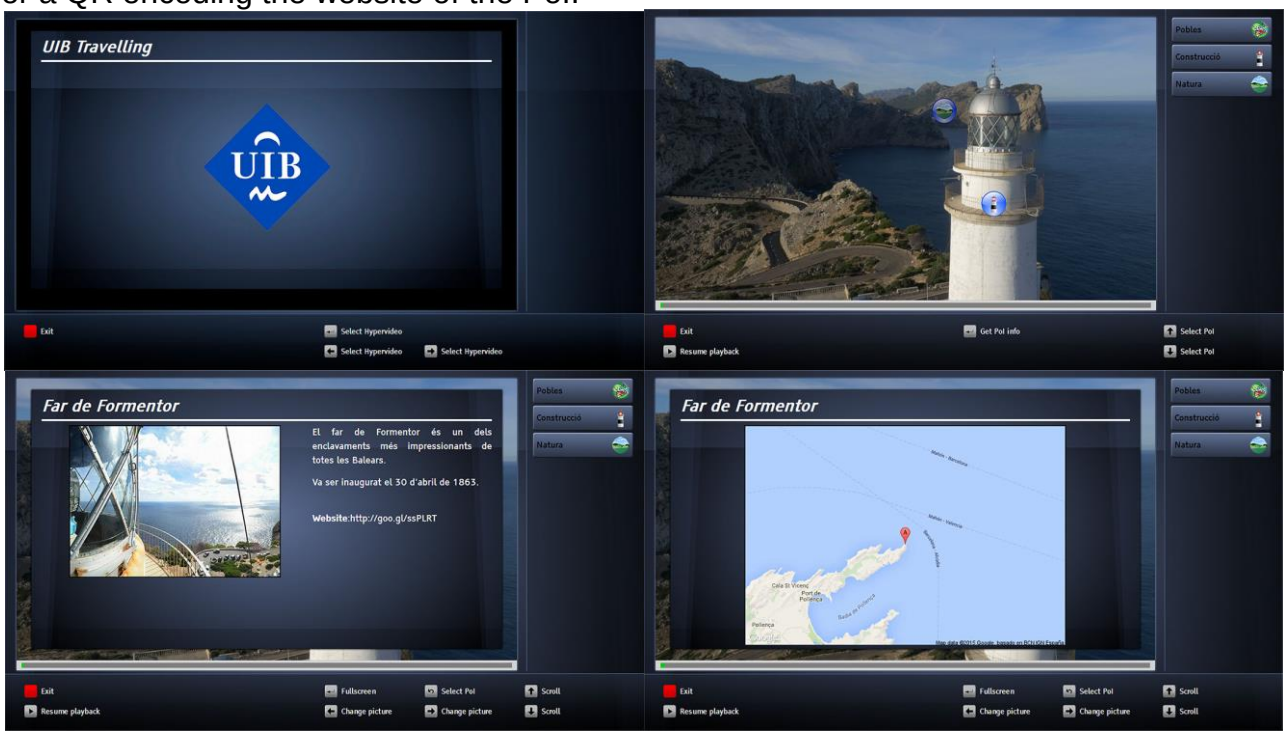


Figure 4. User interface of the visualization module: hypervideo list, hypervideo playback, information window, additional information (more pictures, map, QR).

Finally, a simple design has been followed for the Android second-screen application (Figure 5): after launching the application, the QR scanner opens, to scan the QR shown when the user presses the green button on their remote and link both devices; then a list of the hypervideos appears, synced with the displayed on TV; when a hypervideo is displayed, all the points of interest are shown in a list, plus the currently visible in the TV are highlighted; when the user selects a PoI, they get the additional information in their own device, enabling they to share it to the main screen, seek the multimedia content to the moment(s) when the PoI appears, navigate to the PoI website and open the Google Maps app to the location of the PoI.

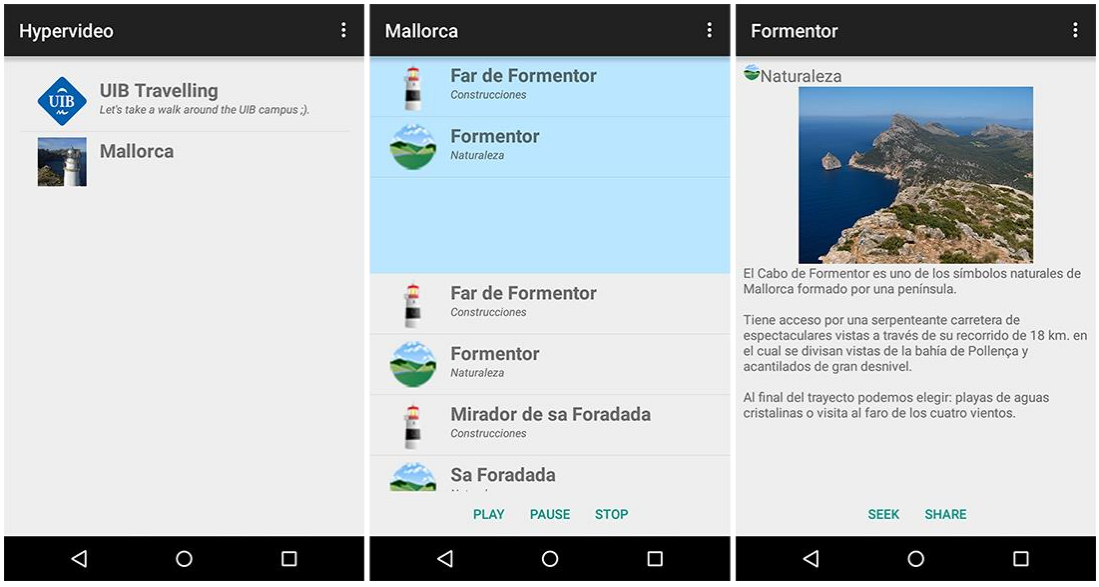


Figure 5. User interface of the second-screen application: hypervideo list, PoI list, PoI window.

Creation module: heuristic evaluation

Design of the evaluation

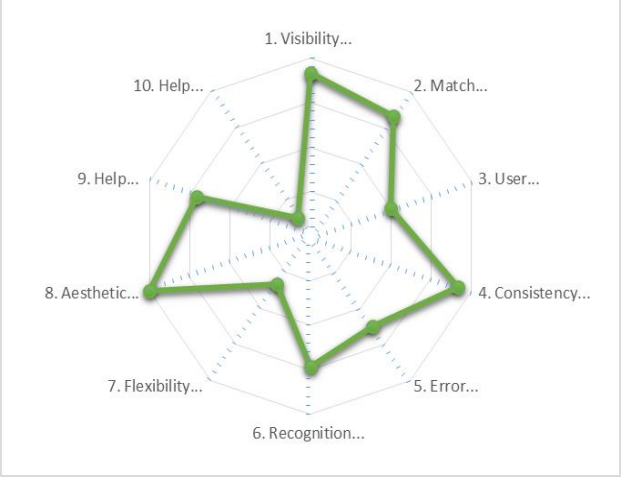
To assess the usability of the creation module, a heuristic evaluation was conducted, following Jakob Nielsen's heuristics [7]:

1. Visibility of System Status
2. Match Between System and the Real World
3. User Control and Freedom
4. Consistency and Standards
5. Error Prevention
6. Recognition Rather Than Recall
7. Flexibility and Minimalist Design
8. Aesthetic and Minimalist Design
9. Help Users Recognize, Diagnose, and Recover From Errors
10. Help and Documentation

To do so, Deniese Pierotti's checklist [8] was applied to every screen, form and menu of the creation module, assigning the punctuation to each heuristic according to the ratio between positive answers and positive + negative answers.

Results of the evaluation

High values (Figure 6) were scored in some heuristics, such as in 1. Visibility of System Status, 2. Match Between System and the Real World, 4. Consistency and Standards or 8. Aesthetic and Minimalist Design, but lower values in heuristics like 7. Flexibility and Minimalist Desgin and 10. Help and Documentation show usability problems that needs to be adressed.



Heuristic	Value
1.	91%
2.	83%
3.	50%
4.	92%
5.	63%
6.	73%
7.	33%
8.	100%
9.	71%
10.	13%

Figure 6. Results of the heuristic evaluation

For each heuristic, these actions are proposed to improve the usability of the system: (1) provide feedback indicating what to do next when an action is completed; (2) properly format numeric values; (3) implement undo and repeat actions; (4) provide a proper description in input fields; (5) provide default values in input fields and show the number of letters left to write; (6) show which input fields are optional and move to the first place the most important word in menus; (9) rewrite error messages, providing their severity and how to solve them; (10) provide a help menu.

Visualization module: experiment between users

Design of the experiment

In order to analyze the usability of the visualization module and compare both solutions, the previous TV-only system and the alternative second-screen approach, an experiment between users was performed.

Ten users were asked to carry out the experiment, consisting in executing four tasks with the TV-only system (system A onwards) and five tasks with the alternative second-screen system (system B onwards) in a *think-aloud* fashion. The order of systems tested was changed user to user to eliminate any learning effect. The null hypothesis is defined as "there is no difference in usability when using system A or system B".

The tasks executed were the following:

- System A (TV-only):
    2. Select and playback the hypervideo "Mallorca"
    3. Obtain additional information from PoI "Sa Foradada"
    4. Check the location of this PoI
    5. Visit the website of this PoI with your mobile device
- System B (second-screen):
    1. Link your second-screen device with the TV
    2. Select and playback the hypervideo "Mallorca"
    3. Obtain additional information from PoI "Sa Foradada"
    4. Check the location of this PoI in the Google Maps app
    5. Visit the website of this PoI

Tasks 2-5 match between systems, in order to be able to compare both solutions.

During the experiment, the following values of efficiency and effectivity, commonly used in TV usability tests, were measured for every task and user:

- Time consumed to complete a task
- Number of button presses to complete a task
- Number of failed attempts to complete a task
- Number of times a user asked help to complete a task
- Number of tasks completed

After testing each system, the users answered the System Usability Scale questionnaire [9], widely used in TV usability tests, to obtain a measure of the user satisfaction with the systems. SUS has been proven to be a reliable and valid measure to assess the usability of a system in a fast and low-cost way [10].

The experiment was performed using the HbbTV broadcast signal on a 40 inch Samsung SMART TV from year 2012, at 2.5 meters. The second-screen device used was a LG G3 smartphone with a 5.5 inch Quad-HD display, with Android Lollipop OS.

The user sat on a couch, with a table in front of it, trying to emulate a dining room as much as possible in the laboratory.

Results of the experiment

The mean time to complete the tasks is shown in Figure 7, highlighting a large difference in the time consumed by users when performing tasks 2-5, rendering system B a better solution.

Task 3 in system A was the most time-consuming because it was a time-dependent task: the user had to wait for the point of interest to appear in the video and pause the media in that specific time. Many users didn't achieve that and had to repeat the task, resulting in an increased time. While on system B all times are low, task 1 presents a pretty high time, due to a lack of an explanation on how to link both devices.

The number of clicks needed to perform each task also benefits System B. An explanation of this results lies in the sequential access with a remote controller versus the direct access with a smartphone touch screen.

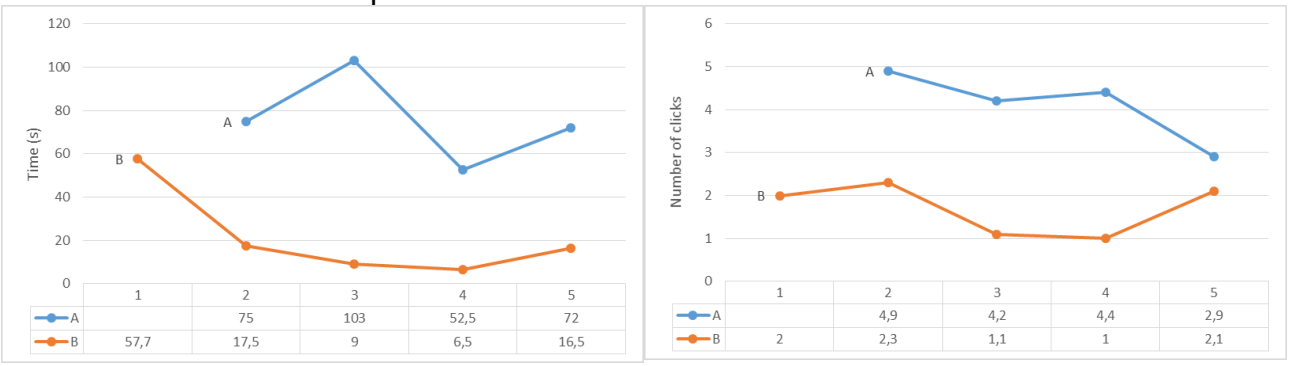


Figure 7. Mean time & mean clicks to complete a task

In Figure 8 the mean number of errors is compared between systems. An error is counted when the user takes a wrong path. 7 out of 10 users did not know that they had to pause the media in task 3 for system A, resulting in a high error count. With system B, users had problems with tasks 1 and 5. For the first one, a lack of an explanation on how to link both devices led to an increase in errors, while for the last one, 4 out of 10 users tried to access the website of the point of interest through the Google Maps interface (they are there as a result from the previous task).

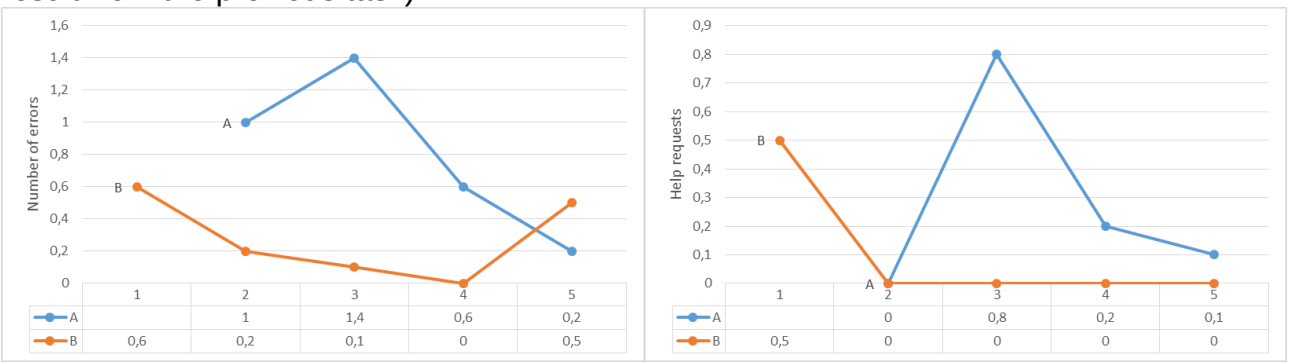


Figure 8. Mean errors & mean help requests to complete a task

Finally, regarding the number of help requests, as can be deduced from previous measures, task 3 for system A and task 1 for system B had a high number of help requests. For system A, the help requests in tasks 4 and 5 were due to the lack of an indication of how to access the map and the QR code.

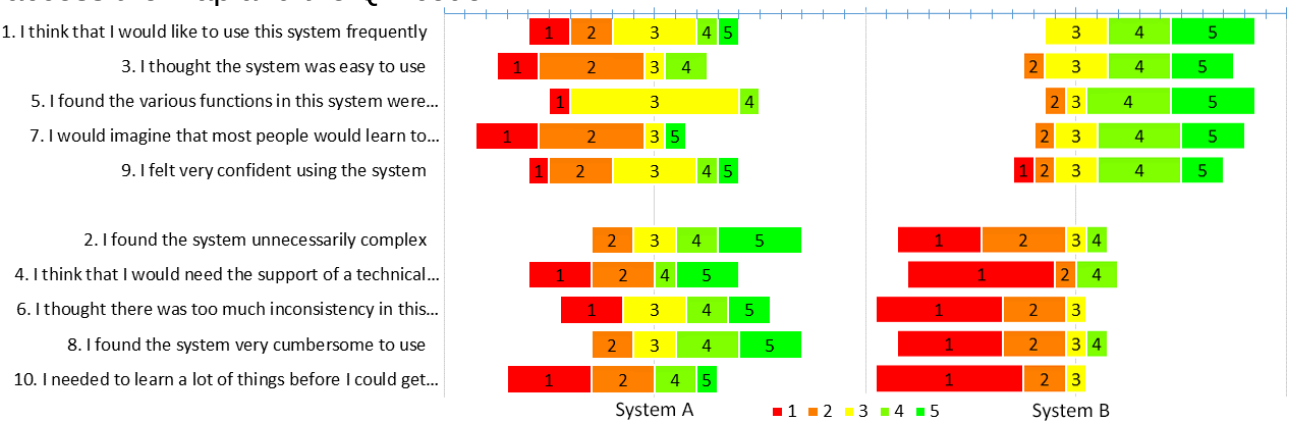


Figure 9. System Usability Scale (SUS) results for Systems A & B

System A scored 43 in SUS, with a SD of 17.3, while system B scored 77.8, with a SD of 16.52. Individual values for each questio–n can be seen in Figure 9.

Bangor et al. (2009) consider that systems with scores less than 70 should be considered candidates for continued improvement [11], so therefore, System B would be considered in the Acceptable range, but System A should be enhanced.

With these results, the introduction of a second-screen application seems to improve the overall usability of the visualization module. To ensure that the improvement observed is caused by the use of this new system a paired t-test and an ANOVA test were ran.

		Time	Clicks	Errors	Help
T2	p-value	0.009353	0.1748	0.01071	NA
	95% IC	17.98413 97.01587	-1.392919 6.592919	0.2357189 1.3642811	NaN NaN
	Mean diff	57.5	2.6	0.8	0
T3	p-value	0.006801	5e-04	0.01328	0.01071
	95% IC	33.11757 154.88243	1.774497 4.425503	0.3432148 2.2567852	0.2357189 1.3642811
	Mean diff	94	3.1	1.3	0.8
T4	p-value	0.002721	0.001587	0.02386	0.1679
	95% IC	20.55143 71.44857	1.673895 5.126105	0.09981823 1.10018177	-0.101621 0.501621
	Mean diff	46	3.4	0.6	0.2
T5	p-value	0.002643	0.327	0.08113	0.3434
	95% IC	24.94227 86.05773	-0.9457575 2.5457575	-0.64555021 0.04555021	-0.1262157 0.3262157
	Mean diff	55.5	0.8	-0.3	0.1

Figure 10. Paired t-test results

Figure 10 contains the results of applying a paired t-test to the data gathered from the experiment, comparing both systems. Cells with p-value < 0.05 represent a significative difference between both systems. It can be observed that the difference in time is signicant in all tasks, while other measures are not significant in every task.

Time	SS	num Df	Error SS	den Df	F	Pr(>F)	
(Intercept)	154880	1	20008	9	696.699	1,57E-02	***
System	80011	1	19264	9	373.812	0.0001762	***
Task	7188	3	38600	27	16.758	0.1956779	
System:Task	6681	3	40969	27	14.677	0.2454695	
Clicks	SS	num Df	Error SS	den Df	F	Pr(>F)	
(Intercept)	655.51	1	29.612	9	1.992.271	1,91E-04	***
System	122.51	1	30.112	9	366.164	0.0001902	***
Task	14.94	3	215.437	27	0.6240	0.6056575	
System:Task	20.34	3	178.537	27	10.252	0.3970002	
Errors	SS	num Df	Error SS	den Df	F	Pr(>F)	
(Intercept)	20.0	1	3.5	9	514.286	5,24E-02	***
System	7.2	1	2.3	9	281.739	0.0004887	***
Task	2.7	3	13.8	27	17.609	0.1784054	
System:Task	6.7	3	11.8	27	51.102	0.0062634	**
Help	SS	num Df	Error SS	den Df	F	Pr(>F)	
(Intercept)	15.125	1	13.625	9	99.908	0.011535	*
System	15.125	1	13.625	9	99.908	0.011535	*
Task	19.375	3	26.875	27	64.884	0.001893	**
System:Task	19.375	3	26.875	27	64.884	0.001893	**

Figure 11. Two-way repeated measures ANOVA test results

We also performed a two-way repeated measures ANOVA, as two independent variables were compared (system and task). Its results are shown in Figure 11.

For task completion time, the effect of system was statistically significant ($F_{1,9}$ = 373.812, p < .001), but the effect of task was not statistically significant ($F_{3,27}$ = 16.758, ns). The effect of the interaction system:task was not statistically significant either ($F_{1,9}$ = 14.677, ns).

For clicks to complete a task, the effect of system was statistically significant ($F_{1,9}$ = 366.164, p < .001), but the effect of task was not statistically significant ($F_{3,27}$ = 0.6240, ns). The effect of the interaction system:task was not statistically significant either ($F_{1,9}$ = 10.252, ns).

For number of errors to complete a task, the effect of system was statistically significant ($F_{1,9}$ = 281.739, p < .001), but the effect of task was not statistically significant ($F_{3,27}$ = 17.609, ns). The effect of the interaction system:task was also statistically significant ($F_{1,9}$ = 51.102, p < .01).

For task completion time, the effect of system was statistically significant ($F_{1,9}$ = 99.908, p < .05), the effect of task was also statistically significant ($F_{3,27}$ = 64.884, p < .01), and the effect of the interaction system:task was also statistically significant ($F_{1,9}$ = 64.884, p < .01).

Upon the completion of the experiment, the key points that had to be improved in the TV application are the selection of a point of interest, the access to its map and QR code. Many users misunderstood the category chooser to select a specific point of interest, resulting in hiding the entire category. For this reason, this menu will be removed in a future version. In the PoI information window a reference to all the media available (pictures, map & QR) will be displayed as thumbnails, showing the user that they can find more information there. A suggestion of the users that will be taken into account is to reduce the number of remote keys required to operate the application.

On the other side, a detailed help window needs to be introduced before starting the linking process the mobile application, telling them to press the required button in the TV application. Finally, a clearer way of telling which points of interest are currently visible has to be found.

Finally, it is observed that the second-screen application provides a more usable experience than the traditional TV and remote controller, and this difference is caused by the differences in the systems, according to the significance of the ANOVA and t-tests performed, thus rejecting the null hypothesis.

Conclusion

Thanks to performing this study, the usability of the entire Hypervideo platform has been tested: both in the creation module, with a heuristic evaluation, and in the visualization module, with a between-users experiment, revealing some usability problems not previously detected.

The heuristic study of the creation module allowed us to identify some aspects not taken into consideration that could significantly improve the users' experience with the system, helping them to have an order to follow in the application.

Then, thanks to the user experiment, first-hand user opinion has been obtained, enabling us to understand which parts of the system confuse the users and why. We also could assert that, although it needs a little tuning, the second-screen application does actually improve the usability of the system.

Finally, when the identified improvements detected are implemented, the Hypervideo platform would be a final product, ready for the use of the general public.

Acknowledgements

This work was supported by project ConTVLab IPT-2012-0871-430000 of the Spanish Government and by the Red AUTI 512RT0461 granted by the CYTED Programa Iberoamericano de ciencia y tecnología para el desarrollo and TIN12-35427 granted by the Spanish Government.

References

[1] Shawhney, N., Balcom, D., Smith, I. Hypercafe: Narrative and Aesthetic Properties of Hypervideo. In Proc. Hypertext 1996, ACM (1996), 1-10.

[2] Landow, G., Kahn, P. Where's the Hypertext? The Dickens Web as a System-Independent Hypertext. In *Proc. Hypertext* 1992, ACM (1992), 149-160.

[3] Bibiloni, T., Mascaró, M., Palmer, P., Oliver, A. Realidad Aumentada en HbbTV: Implementación de una plataforma Hypervideo para la Televisión Digital Conectada. In Proc. CISTI 2014, AISTI (2014), 743-748.

[4] Bibiloni, T., Mascaró, M., Palmer, P., Oliver, A. Hypervideo: Augmented Reality on Interactive TV. In Proc. jAUTI 2014, ACM (2014).

[5] Bibiloni, T., Oliver, A. Augmented Reality on HbbTV: An Hypervideo approach. In Demo Proc. TVX 2014, ACM (2014).

[6] Bibiloni, T., Mascaró, M., Palmer, P., Oliver, A. Hypervideo: Augmented Reality on Interactive TV. In Proc. TVX 2015, ACM (2015).

[7] Nielsen, J., 10 Usability Heuristics for User Interface Design, Conf. companion Hum. factors Comput. Syst. CHI 94, pp. 152–158, 2005.

[8] Pierotti, D., Heuristic evaluation-a system checklist. Xerox Corporation, 1995.

[9] Brooke, J., SUS-A quick and dirty usability scale. Usability evaluation in industry, pp. 189–194, 1996.

[10] Sauro, J., A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices. Denver, CO: Measuring Usability LLC, 2011.

[11] Bangor, A., Kortum, P., Miller, J. A., The System Usability Scale (SUS): An empirical evaluation. International Journal of Human-Computer Interaction, 24, pp 574–594, 2008.