A multimodal interaction system for big displays

Ana M. Bernardos, Ana Muñoz, Luca Bergesio, Juan A. Besada, José R. Casar
Information Processing and Telecommunications Center,
Universidad Politécnica de Madrid
Madrid, Spain
anamaria.bernardos@upm.es

ABSTRACT
Big displays and ultrawalls are increasingly present in nowadays environments (e.g. in city spaces, buildings, transportation means, teaching rooms, operation rooms, convention centres, etc.), at the same time that they are widely used as tools for collaborative work, monitoring and control in many other contexts. How to enhance interaction with big displays to make it more natural and fluent is still an open challenge. This paper presents a system for multimodal interaction based on pointing and speech recognition. The system makes possible for the user to control the big display through a combination of pointing gestures and a set of control commands built on a predefined vocabulary. The system is already prototyped and being used for service demonstrations for different applications.

Author Keywords
Big displays, ultrawalls, multimodal interaction, pointing system, speech interaction, natural interfaces, user experience.

ACM Classification Keywords
H.5.2 User Interfaces, D.2.2 Design Tools and Techniques, H.1.2 User/Machine Systems.

INTRODUCTION
Big displays are usually deployed to deliver quasi-immersive experiences, to facilitate the consumption of different information channels in an opportunistic way, to visualize and manage large amount of data or to enrich collaborative working contexts. Their big size format enables to show all the information at one glance, avoiding exhausting screen shifts that may be needed when using smaller formats.

How to achieve fluid interaction with big displays is still a challenge. Touch interfaces – directly enabled through the big wall's tiles or through support devices (e.g. mobile devices, tablets or laptops) are common. When the display surface is touch-sensitive, the user has to approach the screens up to a short distance, missing the general view of it and needing to move intensively. Regarding device-assisted approaches, they require that the user continuously handles an instrument, which may come to be limiting in some interaction workflows. Previous works (e.g. [1]) have explored how to build a sound pointing system over a big display in mid-distances. Some other studies have proposed the use of speech and gesture recognition to deliver a solution to perform control and monitoring, in specific scenarios (e.g. crisis management [2]).

This paper presents a multimodal interaction system for big displays that facilitates handling the information in the display through a non-instrumented method, by defining an interaction workflow that combines pointing gestures [3] and a set of speech commands. The system has been prototyped and applied to a control-like scenario for drones' fleet control.

DESCRIPTION OF THE INTERACTION METHOD
We consider that the presentation layer of a big display is a composition of different tiles. The tile is the minimum manageable unit in the wall; it may be screen-size or smaller, or expand over several screens. Each tile is configured to handle a single channel or information pipe. The information pipe may contain multimedia content, raw data, maps or communication applications, for example, so different sets of actions have to be available.

We also assume that, in general, when managing information channels in a big display, it is useful to count on a set of user-predefined layouts to package the tiles (Figure 1), with the purpose of giving more relevance to specific information channels if needed. The multimodal interaction method facilitates to: a) select tiles of information by pointing at them; b) perform actions over the selected tiles through speech commands and c) use speech commands to swap between different layouts.

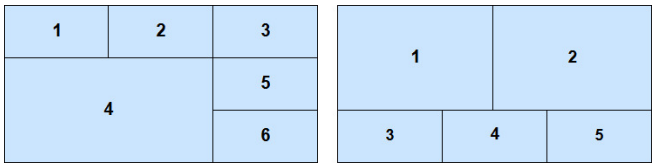

Figure 1. Example of two predefined layouts in a 9-screen ultrawall. Configuration a) focuses on tile 4, while configuration b) balances the information of tiles 1 and 2.

PerDis '17. June 07-09, 2017. Lugano, Switzerland.

ACM 978-1-4503-5045-7/17/06.
http://dx.doi.org/10.1145/3078810.3084353

The interaction workflow initiates from a default layout in which all the available information channels occupy the same surface in the display (i.e. the display is divided into equally-sized tiles, the number of tiles depends on the number of information channels). On this, the user may directly switch to another layout through a speech command. The designed system contains a set of predefined layouts, each of them stored with a name the user must know. When the user wants to swap the layout, he must say the command *"change to interface"* followed by the name of the new interface. This action can be done at any moment.

As an alternative, the user may choose to interact with a single tile in the layout by pointing at it and then using a command configured in the control vocabulary (Table 1, *tile interaction*). The vocabulary has been designed taking as a reference the results of a user experience study that was carried out in November 2016. In that study, 17 people chose their preferred gestures and spoken commands for different actions over the big display on a specific service scenario.

Type of interaction	**Available commands**
Layout mngmt.	"Change interface to"
Tile interaction	*Call*:"Call", "Answer", "Hang up"
	Video: "Play", Pause", "Forward", "Backward"
	General: "Full screen", "Restore", "Expand", "Next", "Zoom in", "Zoom out"
	Maps: "Sign PoI"

Table 1. Available commands for each of the two set of grammars.

Through this simple combination of layout management and tile control, the user can handle the whole information service.

SYSTEM ARCHITECTURE

The system is built on Kinect sensor, which is used to enable pointing and speech recognition. The pointing system is an in-house development [3], while the system uses Kinect APIs for speech recognition. The architecture of the system (Figure 2) is then composed of these two independent modules. The pointing module is composed by:

- *Pointing detection trigger*, configured to initialize the pointing calculation every time the Kinect sensor detects a body.
- *Pointing calculation* module, detects the finger of the user to determine the direction the user is pointing at.
- S*creen selector module,* which decides at which display is the user pointing based on the previous pointing calculations.

On the other hand, the speech recognition module is composed of two elements:

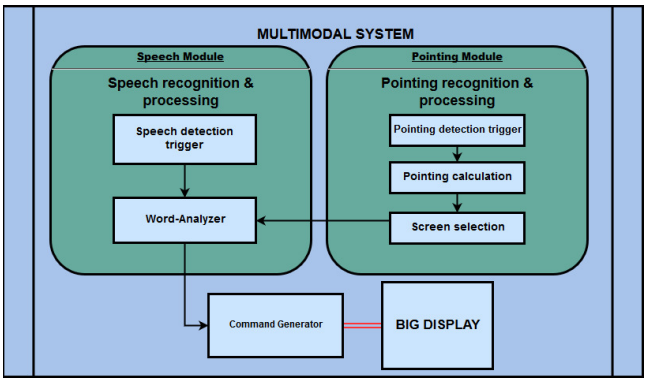

Figure 2. Diagram of the system architecture.

- The *speech detection trigger* is configured to initialize the speech analysis every time the Kinect microphone detects a speech.
- The *word-analyzer module* searches for similarities between the speech and the set of predefined words that compose the vocabulary of this system. Once the module distinguishes whether the user aims at performing a layout change or executing a command, it is also in charge of sending the corresponding command to the server controlling the displays.

In the video annexed to this paper, the multimodal interaction system has been used to prototype a control center for drone fleets, thus enabling the controller to check the real-time data coming from different flying unmanned aerial vehicles.

CONCLUSION

The multimodal interaction system for big displays is still in its prototyping stage, but it already integrates the speech-pointing interaction workflow that enables to complete full management of the information channels of a specific service setting. Following we will be studying how to provide better cues to the user, at the same time that the system functionalities are extended to increase the configuration possibilities. An interesting aspect to explore is how multiple users may manage the big wall in an efficient and coordinated way towards the same objective.

ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under grant TEC2014-55146-R and by UPM under grant RP150955017.

REFERENCES

1. M. Nancel, E. Pietriga, O. Chapuis, M. Beaudouin-Lafon. 2015. Mid-air pointing on ultra-walls. *ACM Trans. Comput.-Hum. Interact*. 22, 5, Article 21, 62 pp.
2. R. Sharma, M. Yeasin et al. 2003. Speech-Gesture driven multimodal interfaces for crisis management. *Proceedings of the IEEE, vol. 91, no. 9*, 1327-1354.
3. A.M. Bernardos, D. Gómez, J.R. Casar. 2016. A Comparison of Head Pose and Deictic Pointing Interaction Methods for Smart Environments. *Intl. J. of Human-Computer Interaction*, *32*(4), 325-