Social Documentary: An interactive and evolutive installation to explore crowd-sourced media content

Fabien Grisard
UMONS
Mons, Belgium
fabien.grisard@umons.ac.be

Ceren Kayalar
Sabanci University
Istanbul, Turkey
ckayalar@sabanciuniv.edu

Sema Alaçam
Istanbul Technical University
Istanbul, Turkey
alacams@itu.edu.tr

Özgün Balaban
Singapore University of
Design and Technology
ozgunbalaban@gmail.com

Yekta İpek
Istanbul Technical University
Istanbul, Turkey
yektaipek@gmail.com

Stéphane Dupont
UMONS
Mons, Belgium
stephane.dupont@umons.ac.be

ABSTRACT

This paper aims to present a project in progress, an interactive installation for collaborative manipulation of multimedia content. The proposed setup consists in a vertical main screen and a horizontal second screen, which is used as control panel, reproducing an augmented physical desktop. Augmented reality markers are used to give the user an intentional way to interact with the system and a depth camera is used to estimate the users' gaze and quantify how interested they are in the displayed content, slightly modifying the video projection itself.

Author Keywords

Interactive installation; Fiducials; Gaze tracking; Attention estimation

INTRODUCTION

With the development and the diffusion of new technologies, many events (social, politics, sport, etc.) involve the production of huge amount of multimedia content. Thus, using them as raw material in artistic works is becoming more challenging but more interesting. The installation we propose is a way to display and to interact with big amount of content, mixing videos from different sources, images and texts. The visitor acts in two ways on the content : by choosing explicitly notions related to the video to display and in a much more unconscious way, by being (or not) interested in it.

RELATED WORKS

Augmented reality marker, also called fiducials, lead to the rapid development of new possibilities in table-top augmented tangible user interfaces. The web is full of examples of use of tangible interfaces in several domains such as artistic and musical creation[1] or educative games[2]. Usually, the artistic installations don't deal with the visitor attention. We want to use it along with the fiducials, as an input of our system. A fiducial is a geometric unique two-dimensional figure which provides presence, orientation, location and identity information in real-time, when placed in the field of view of a camera.

INSTALLATION DESCRIPTION

When the visitors enter the installation, they see a big screen (main screen) on the wall and a table on which a map is projected. There are also three sets of colored objects near the table, corresponding to three classes of keywords: people, action, emotion. Figure 1 presents an illustration of the installation with the detailed components. The visitors select one object of each color and put them on the map. A video segment related to the chosen keywords is displayed on the main screen. When the visitors are looking at the same area, a text is added to the video. Each time a visitor shows interest for a segment of the video, the rating of this segment is incremented, increasing the probability to display it to the next visitors. When all the visitors are gone, the main screen is shut down.

Our videos, images and texts are related to the social and political events that happened at Gezi Park in 2013. During this period, multimedia content have been produced in abundance both by the press and by demonstrators themselves, as mentioned by Alaçam et al.[1]. Videos taken during these events can provide historical documentaries putting together subjective and different views of the same event. Compared to classical historian-made documentaries, this novel approach provides emotion, hope and excitement captured during important moments which make the History.

Video segmentation and annotation

The video collection we found on Facebook mixes images from many sources. Each file is about three-hours-long. We chose not to cut them into short parts but to attach a subtitle-like file which contains useful information about each segment (beginning and end times, viewers interest rate - as Facebook "likes", annotation keywords). Thanks to the European project LinkedTV[3] visual analysis techniques [4], a big part of this work can be automatized. Those techniques include automatic shot segmentation and concepts recognition. The resulting file need to be manually edited to perfectly fit

[1] http://modular-drops.tumblr.com/

[2] http://www.woutersontwerpers.nl/portfolio/bosinfocentrum-t-leen/

[3] http://www.linkedtv.eu/

2nd International Workshop on Interactive Content Consumption at TVX14, June 25, 2014, Newcastle, UK. Copyright is held by the author/owner(s).

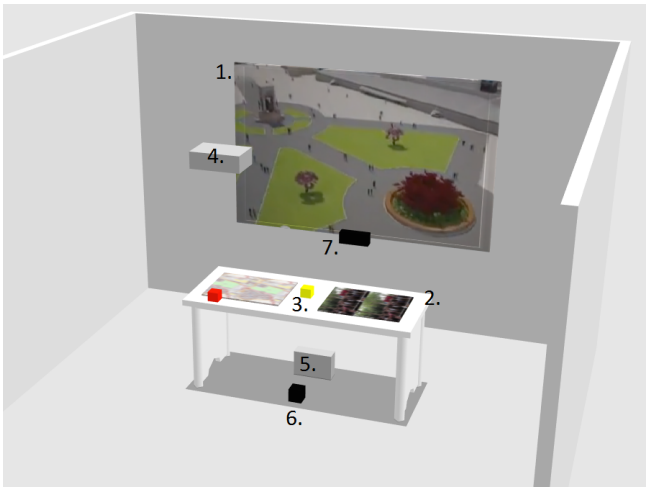

Figure 1. Installation setup with: 1. Main screen, displays the video stream; 2. Second screen, displays a map / Control panel; 3. Augmented reality markers on colored objects; 4. Main video projector; 5. Second projector; 6. webcam for fiducials tracking; 7. MS Kinect sensor for visitors' head tracking

our purpose but, thanks to the subtitle-like format, this part is quite easily done with any subtitle editor. By default, all the segments' rates are set to one.

Attention tracker

As exposed in the introduction, we propose an interactive installation based both on fiducial markers and on the analysis of the visitor attention for the currently played multimedia content. According to Hawkins [2], the time the visitor spends looking at the screen is linked to the attention he has for the projected content. Hawkins defines four "Types of Looks" : *Monitoring (¡= 1.5 seconds), Orienting (1.5-5.5 sec.), Engaged (5.5-15 sec.) and Stares (¿15 sec.)*. To identify which part of the installation the visitor is roughly looking at, we rely on the estimation of his head pose. Accordingly to Murphy-Chutorian [3], "[...] *Head pose estimation is intrinsically linked with visual gaze estimation* [...]. *By itself, head pose provides a coarse indication of the gaze that can be estimated in situations when the eyes of a person are not visible* [...]." The configuration of the installation (the screen width is close to the visitor-screen distance), increases this effect. As a result, if we can estimate the head position and orientation, we can infer on interest of the visitor. To achieve this goal, we use a Kinect sensor and the MS Kinect SDK face tracking functions[4]. Each time a visitor's attention is engaged (duration ¿ 5.5 seconds), the rate of the corresponding segment is incremented.

Segment selection and effects

Any object from wooden or plastic material can be turned into a tangible controller by sticking a fiducial under it. The application displaying the video content is a client listening to the TUIO (UDP based network protocol)[5] messages from the reacTIVision framework[6]. The video content on the vertical projection changes according to the markers put on the table. When the number of objects on the table changes, a list of video segments is updated and contains all the segments annotated with the keywords corresponding to the objects. To avoid rapid convergence toward a reduced list of "popular" segments, the one we display is chosen semi-randomly, accordingly to its rate. Once the segment is chosen, the player jump to the beginning time of this segment.

In case of joint attention (two visitors looking at the same screen), we propose to display related tweets on the concerned screen and to increase the sound when the video shows protesters, as an intensification of the "voice of people".

If the user puts more than one item on the table, our system can relate these items together by the distance between them, give a visual feedback on the table and display a combined content on the video projection.

[4]http://msdn.microsoft.com/en-us/library/jj130970.aspx

[5]http://www.tuio.org/

[6]http://reactivision.sourceforge.net/

CONCLUSIONS

This project aims to provide a tangible desktop application to understand, display and interact with a big collection of multimedia content. The installation should be able to adapt itself to the public thanks to attention evaluation and an automatic video rating system.
In our case, the content is related to very localized social events but the system could be used for other purposes, as in museums. This system brings new opportunities to build subjective crowd-based documentaries.

ACKNOWLEDGMENTS

This work is supported by the Integrated Project LinkedTV (www.linkedTV.eu) funded by the European Commission through the 7th Framework Program (FP7-287911).

REFERENCES

1. Alaçam, S., Ipek, Y., Balaban, Ö., and Kayalar, C. Organising crowd-sourced media content via a tangible desktop application. In *MMM (2)*, C. Gurrin, F. Hopfgartner, W. Hürst, H. D. Johansen, H. Lee, and N. E. O'Connor, Eds., vol. 8326 of *Lecture Notes in Computer Science*, Springer (2014), 1–10.

2. Hawkins, R. P., Pingree, S., Hitchon, J., Radler, B., Gorham, B. W., Kahlor, L., Gilligan, E., Serlin, R. C., Schmidt, T., Kannaovakun, P., and Kolbeins, G. H. What produces television attention and attention style? *Human Communication Research 31*, 1 (Jan. 2005), 162–187.

3. Murphy-Chutorian, E., and Trivedi, M. M. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell. 31*, 4 (Apr. 2009), 607–626.

4. Stein, D., Apostolidis, E., Mezaris, V., Abreu, N. D., and Mller, J. Semiautomatic video analysis for linking television to the web. In *In Proc. FutureTV Workshop* (2012), 1–8.