Grab your Favorite Video Fragment: Interact with a Kinect and Discover Enriched Hypervideo

Vuk Milicic
EURECOM
Biot, France
vuk.milicic@eurecom.fr

José Luis Redondo García
EURECOM
Biot, France
redondo@eurecom.fr

Giuseppe Rizzo
EURECOM
Biot, France
giuseppe.rizzo@eurecom.fr

Raphaël Troncy
EURECOM
Biot, France
raphael.troncy@eurecom.fr

ABSTRACT

In this demonstration, we propose an approach for enriching the user experience when watching television using a second screen device. The user can control the video program being watched using a Kinect and can grab, at any time, a fragment from this video. Then, we perform named entity recognition on the subtitles of this video fragment in order to spot relevant concepts. Entities are used to gather information from the Linked Open Data cloud and to discover what the *vox populi* says about this program. This generates media galleries that enrich the seed video fragments grabbed by the user who can then navigate this enriched content on a second screen device. A showcase of this demo is available at `http://youtu.be/4mSC685AG7k`.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*World Wide Web*; H.5.2 [**User Interfaces**]: Screen design

General Terms

Television, User Interface

Keywords

Connected TV, Interactive Television, Hypervideo, Media Fragment, Linked Data

1. INTRODUCTION

Enriching television content and providing a second screen experience represents a new challenge for broadcasters. With the advent of the Linked Open Data (LOD) cloud and social platforms such as Twitter and Facebook, more and more broadcasters aim to mine information from these streams of heterogeneous data. Meanwhile, research efforts have shown the importance of interlinking video content with Web resources to a television user [2], while Aroyo *et al.* proposed a scenario where users receive personalized metadata for enriching their television experience [1]. In this work, we propose a demonstration where a user, through gestures, can capture fragments of the content being played and discover additional content automatically aggregated using the LOD cloud. The resulting content is pushed on a second screen device, enriched with visual galleries created from the *vox populi* of social platforms.

2. TECHNICAL DETAILS

The television scenario developed for this demo is composed of a main screen, a second screen device (tablet) and a gesture tracker that allows the user to interact with the system. The entire user interface has been developed using modern Web technologies (HTML5, CSS, Javascript), hence both the main screen and the second screen are displayed using common Web browsers anticipating the upcoming HbbTV 2.0 specification. The interaction between the Kinect and the main screen is realized through Kinesis.io[1] and parsed by a so-called Sync-Server which, therefore, propagates the actions to the second screen. The, the second screen collects the enrichment from the Hypervideo Annotation component (Figure 1). In the remainder of this section, we further describe the main components of this architecture.

Main screen has two main objectives: showing the television prorgram in full screen mode and enabling the user to navigate through multiple channels in order to select a video program. For providing an intuitive content search and an easy visualization, a mosaic user interface has been implemented, taking inspiration from [5]. The video items are provided by the LinkedTV[2] streaming server. Kinesis turns this component into a gesture enabled application. The following gestures are recognized: *i)* swipe-up (video play), *ii)* swipe-down (grab shot), *iii)* swipe-right (video forward), *iv)* swipe-left (video backward). Through Javascript calls, the application sends to the Sync-Server the actions that the user has performed.

[1] `http://kinesis-io.github.com`
[2] `http://www.linkedtv.eu`

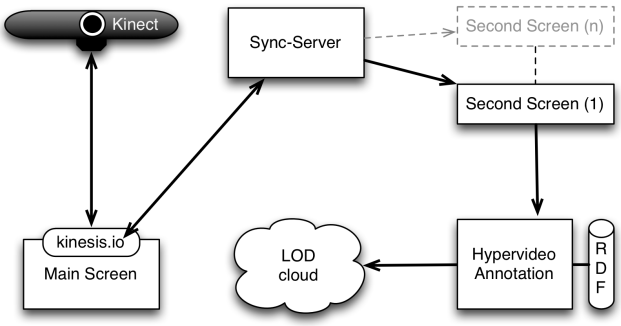

Figure 1: Architecture: enriched video program on a second screen

Sync-Server is the core of the architecture. It holds the synchronization between the gestures captured on the main screen and the user interface rendered on the second screen. It is powered by a NodeJS[3] server and it keeps the connections alive with the second screen(s) through WebSockets. When a recognized gesture (event) occurs on the main screen, this component forwards the corresponding action to the different second screen devices.

Second screen receives from the Sync-Server the actions forwarded by the main screen. A master/slave paradigm controls the harmonization between this component and the main screen. According to the main screen's timestamp and the video identifier received from the Sync-Server, it collects the video fragment annotations available in the Hypervideo Annotation repository. Different second screen devices can be connected to the Sync-Server at the same time.

Hypervideo Annotation repository contains the video metadata. Traditionally, EPGs (Electronic Program Guides) provide only coarse and static information about an entire program. In our approach, rich metadata about the program is instead represented according to the LinkedTV ontology[4]. This model defines a list of classes that are relevant in the television domain such as `Chapter`, `Scene`, `Visual Concept`, `Visual Object` and relies on other well-known ontologies like the Open Annotation Core Data Model[5], the Ontology for Media Resources[6], the NERD ontology, and the Programmes Ontology[7]. Using the Media Fragments URI 1.0 specification, television content can be annotated not only at the level of the entire program but also at different degrees of granularity (in the current demo, chapters, subtitles, and shots). Those individuals are the anchors for attaching information such as legacy metadata from the providers, automatic analysis results like concept detection or face recognition, and entities with links to other resources in the Web where extra information about the content can be found.

For identifying those entities that are relevant to a particular television content, a multilingual entity extraction is performed over the video subtitles by using the NERD framework[4]. The entities spotted are classified using the core NERD Ontology[8] and disambiguated with URIs when possible. Finally, those entities are also used as input for retrieving additional media items on the Web. From a viewer's perspective the content about a particular television program that the crowd is sharing in the Web really matters because it brings interesting insights about what is happening now. After previous research works in this field, we have used the framework MediaFinder [3] for retrieving and analyzing fresh media content. MediaFinder collects microposts shared on social platforms that contain media items as a result of a text query. It automatically creates different visual storyboards that reflect what users have shared about this particular event, leveraging on: *i)* visual features from media items for near-deduplication and *ii)* textual features from status updates to interpret, cluster, and visualize media items. By querying for the label of every extracted entities, MediaFinder returns related items that are also attached to the media fragment a particular entity belongs to. This way, the main seed video program is illustrated with fresh photos and videos that correspond to people feelings and emotions.

3. CONCLUSION AND FUTURE WORK

In this paper, we have presented a framework which exploits the Kinect interaction for allowing a user to grab a video fragment from a main screen where a video is displayed and to splash it on a second screen. Such an action enables the user to discover more pre-processed metadata about the content that has been captured, using both the LOD cloud and numerous social platforms. We aim to build hypervideos that connect streams of heterogeneous data coming from the Web, projected on a second screen where the user ultimately navigate through.

4. REFERENCES

[1] L. Aroyo, L. Nixon, and L. Miller. Notube: The television experience enhanced by online social and semantic data. In *International Conference on Consumer Electronics (ICCE'11)*, Berlin, Germany, 2011.

[2] L. B. Baltussen and J. Oomen. Antiques interactive. In *2nd International Workshop on Personalized access to cultural heritage (PATCH'12)*, Nara, Japan, 2012.

[3] G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J. L. Redondo García, and R. van de Walle. What Fresh Media Are You Looking For? Retrieving Media Items from Multiple Social Networks. In *International Workshop on Socially-aware multimedia (SAM'12)*, Nara, Japan, 2012.

[4] G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for computational Linguistics (EACL'12)*, Avignon, France, 2012.

[5] A. Veenhuizen, R. van Brandenburg, and O. Niamut. MosaicUI: Interactive media navigation using grid-based video. In *10th European Interactive TV Conference (EuroITV'12), Poster track*, Berlin, Germany, 2012.

[3] `http://nodejs.org/api`
[4] `http://semantics.eurecom.fr/linkedtv`
[5] `http://www.openannotation.org/spec/core`
[6] `http://www.w3.org/ns/ma-ont`
[7] `http://purl.org/ontology/po`
[8] `http://nerd.eurecom.fr/ontology/nerd-v0.5.n3`