FaceCloud: Heterogeneous Cloud Visualization of Multiplex Networks for Multimedia Archive Exploration

Benjamin Renoust
National Institute of Informatics & JFLI CNRS UMI 3527, Tokyo JP
renoust@nii.ac.jp

Haolin Ren
National Audiovisual Institute (INA) & University of Bordeaux, CNRS UMR 4800, Bordeaux FR
haolin.ren@labri.fr

Guy Melançon
University of Bordeaux, CNRS UMR 4800, Bordeaux FR
guy.melancon@labri.fr

Marie-Luce Viaud
National Audiovisual Institute (INA), Paris FR
mlviaud@ina.fr

Shin'ichi Satoh
National Institute of Informatics, Tokyo JP
satoh@nii.ac.jp

ABSTRACT

Multimedia data is by nature heterogeneous, conveying semantic information through multiple cues. Text analysis of closed captions already brought us understanding of the spoken information. Today's advances in computer vision now enable us to look for relevant semantic information from the visual content of real-world archives. Combining these two levels of extracted information to make sense of an archive still remains a challenge. Multiplex networks, which model multiple families of interactions in a graph, can capture and combine both sources of semantics. We can leverage on these objects to extract hierarchies and integrate them in an interactive heterogeneous "visual cloud". Inspired by word clouds, these clouds allow to grasp visual and textual semantic information captured from a multimedia collection all at once. The interaction then enables direct access to the relevant video. We demonstrate our system with the exploration of a Japanese news archive.

KEYWORDS

heterogeneous data; visual analytics; media analytics; multiplex networks; tag cloud; NHK; Japan

1 INTRODUCTION

Multimedia analytics, or visual analytics of multimedia data [2], aims at providing high-level representation of multimedia data and collections for the purpose of high-end analysis. It can serve the purpose of helping public figures monitor their public image, of monitoring media impact,of predicting market behavior, or just help users better access relevant information and explore a large multimedia archive.

To that mean, we often need a semantic level of annotation of videos (or parts of videos) that describes the multimedia document. Visual analytics has often tackled this task by relying on text analysis – when text is available from the multimedia data [4]. Recent computer vision tools now provide reliable enough annotation means that allow us to build upon. Another challenge then raises from the visual presentation of this heterogeneous information, which combines both textual and visual semantic concepts.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10.
https://doi.org/10.1145/3123266.3127917

We propose to use multiplex networks in order to model proximity of query results in a multimedia database. These networks allow the combination of all sorts of semantic concepts, while offering to extract hierarchical relationships between them. From this hierarchy of concepts, we construct a hybrid "visual cloud" combining image and textual information extracted from the query results – that we call FaceCloud (Fig. 1).

Figure 1: Presentation of FaceCloud search results interface, videos queried from Shinzo Abe's face.

We use FaceCloud (Fig. 1) to explore queries on a database of news archived from NHK (the Japanese public broadcast company) between 2001 and 2013. The database is composed of 30mn-long news broadcasts, in which we extracted news segments, face detection and tracking of 139 individuals, and topical extraction for each news segments. For each query on the database, we generate an interactive FaceCloud combining text and extracted images, which allows users to access to individual segments of video and refine their search on interaction.

2 DESCRIPTION OF THE SYSTEM

A news program usually presents different news stories that we refer to news segments. For each news program, segments boundaries are detected through a sliding window of topics as provided by the authors of [3]. For each detected segment (Fig. 2), we further extract relevant (Japanese) topics with the GooLabs API then translated with Ms. Bing Translator for convenience.

We then use our large-scale face detection and tracking framework [6] (Fig. 2). It detects faces in each frame using Viola-Jones[12] and regroups instances through point tracking [10]. Face-tracks are then sampled in a few face instances, for each we extract the OpenFace 128-dim embedding[1], averaged among samples. We cluster face-tracks with GreedyRSC [5] and manually annotated 3k+ clusters covering about 15k face-tracks in total and 139 unique public figures. Our database indexes for each video: news segments (with captions), topics, and face-tracks. A query in this database can be placed on three criteria: time-frame, faces, and keywords. Each query returns a subset of documents (news segments) with their associated semantic concepts (detected faces and topics).

Similarly to Detangler [7], we construct a multiplex network, in which nodes are news segments, and one layer of edges for each semantic concept shared by at least two documents. From this network, we construct the layer interaction network [8], which is a weighted network of co-occurring layers on edges (Fig. 3). In this network, a layer becomes a node associated to a computed entanglement index [8], which translates how much a mixes with others. Using this weighted layer interaction network, we proceed to a greedy hierarchical clustering and create a tree of semantic concepts [9].

From this point on, we can use the formed hierarchy to create a "visual cloud" [9] (Fig. 4). We first draw the tree as a modified *pack* layout [13], in which leaf nodes are also assigned a circle. We use then use the *pack* layout positions to initialize a modified *Wordle* tag-cloud algorithm [11]. In order to take images into account, we extended the algorithm to all 2D shapes and constrained spiral displacement of each "tag" to the *pack* layout circles.

The size of each tag and images is mapped to its occurrence, and its color to the group it belongs to. Because we have a hierarchical clustering, we can change the number of groups without having to recompute the layout – it is only a cut in a tree, hence the visual representation remains stable. Mouse hovering highlights the neighbors of a concept in the layer network. We also provide a heatmap-inspired coloring (Fig. 4) to highlight the most entangled layers in the network (which are not necessarily the most occurring one). While hovering, the entanglement corresponds to the subgraph induced by the hovered concept [7].

Query results are presented similarly to any search engine query. A click on a concept in the FaceCloud filters the query results. A click on a query result loads the news video while positioning it at the beginning of the segment. We additionally propose a timelime exploration to locate the resulting documents in the archive, as well as to filter the query results from a selection.

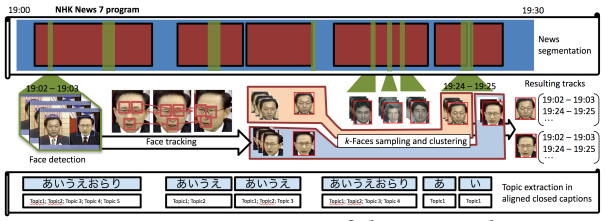

Figure 2: Preprocessing of the news videos.

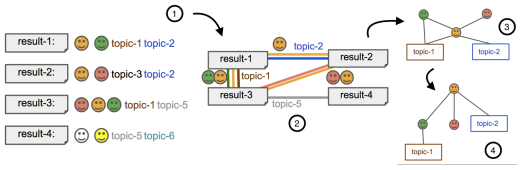

Figure 3: Hierarchy extraction from the search results. 1: indexed video segments. 2: multiplex network of results. 3: associated layer-interaction network. 4: derived hierarchy.

Figure 4: Construction of the visualization. *Pack* layout (left), FaceCloud (center), interactive heatmap (right).

3 CONCLUSION

This system allows to interactively explore the database and to present the content of a search result in a novel way. For example, by searching for current Prime Minister A. Shinzo's face detection occurrences (Fig. 1), filtering results during Prime Minister J. Koizumi's terms, we can see at a glance the different stories that put him in front the politico-media scene were all related to the crisis management of abduction of Japanese citizens by North Korea.

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. Technical report, CMU-CS-16-118, CMU School of Computer Science.

[2] Nancy A Chinchor, James J Thomas, Pak Chung Wong, Michael G Christel, and William Ribarsky. 2010. Multimedia analysis+ visual analytics= multimedia analytics. *IEEE computer graphics and applications* 30, 5 (2010), 52–60.

[3] Ichiro Ide *et al.* 2004. Topic threading for structuring a large-scale news video archive. *Image and Video Retrieval* 1, 1 (2004), 123–131.

[4] Kuno Kurzhals, Markus John, Florian Heimerl, Paul Kuznecov, and Daniel Weiskopf. 2016. Visual Movie Analytics. *IEEE Transactions on Multimedia* 18, 11 (2016), 2149–2160.

[5] D. D. Le and S. Satoh. 2011. Indexing Faces in Broadcast News Video Archives. In *2011 IEEE 11th ICDM Workshops*. 519–526.

[6] Thanh Duc Ngo *et al.* 2013. Face retrieval in large-scale news video datasets. *IEICE Trans. on Information and Systems* 96, 8 (2013), 1811–1825.

[7] Benjamin Renoust, Guy Melancon, and Tamara Munzner. 2015. Detangler: Visual Analytics for Multiplex Networks. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 321–330.

[8] Benjamin Renoust, Guy Melançon, and Marie-Luce Viaud. 2014. Entanglement in multiplex networks: understanding group cohesion in homophily networks. In *Social Network Analysis-Community Detection and Evolution*. Springer, 89–117.

[9] Benjamin Renoust, Haolin Ren, Guy Melançon, and Marie-Luce Viaud. 2017. Layer Hierarchization in Multiplex Networks for Word Cloud Visualization. In *Visualization Symposium (PacificVis), 2017 IEEE Pacific*. IEEE.

[10] Jianbo Shi and Carlo Tomasi. 1994. Good features to track. In *Proceedings CVPR'94*. IEEE, 593–600.

[11] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory visualization with Wordle. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 1137–1144.

[12] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

[13] Weixin Wang, Hui Wang, Guozhong Dai, and Hongan Wang. 2006. Visualization of Large Hierarchical Data by Circle Packing. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, NY, USA, 517–520.