Auracle: how are the salient cues situated in audiovisual content?

Principal investigators: Christian Frisson, Nicolas Riche, from the numediart Institute of the University of Mons (Belgium)

19 December 2014

This eNTERFACE14[1] project proposal aims at studying the mutual influence of audio and visual cues on spectators of audiovisual media content: feature films, animation movies, video games recordings by undertaking an eye tracking user experiment on a database of such content. Based on the results, models and algorithms of the relation between sound and image may be refined and their predictions compared with actual gaze fixations recorded from participants. The media and fixations database will be released for free. The results of such analyses may serve to breed new crossover media genres featuring sound and moving images.

[1] `http://aholab.ehu.es/eNTERFACE14/`

Project objectives

We are eager to investigate how sound alters the gaze behavior of people watching moving images. We believe that a collection of audiovisual fragments sourced not only from feature films, but also animation movies and video game in-take scenes, or even video games actually being played, could provide a wide range of coarse to fine material, of various realistic renderings, to better understand the relation between sound and image in such audiovisual content. We plan to run eye tracking tests on such a database with a statistically signifiant number of people, and release the whole publicly for further research and modeling, what would be started by our team throughout the project. Potential research tracks, depending on the participants interests and profiles, include: multimedia summarization, saliency modeling, annotation and aided understanding of artworks.

Background information

Eye tracking

Eye tracking research has been ongoing for several decades, effervescent in the last, discovered a couple of centuries ago. Eye trackers are to reach a plateau of technological acceptance by everyday users in a few years. Some predictions foresee it as the new trend in video game controllers, "hands-free". The price and form-factor size of these devices is plunging, about soon to be part of consumer-grade video game hardware and to be integrated in mobile devices, most probably through infrared pupil tracking [2].

[2] Corey Holland and Oleg Komogortsev. Eye tracking on unmodified common tablets: Challenges and solutions. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 277–280. ACM, 2012. DOI: 10.1145/2168556.2168615

Available databases

While there are plenty of image and video databases [3,4], we think there is a need for a database of audiovisual stimuli and related human fixations to be made available for free and aimed at building and validating models on the relation between image and sound, with diverse genres beyond movies, documentaries and broadcasts, such as animation [5] and video games, focusing on special directing and sound design techniques [6]. It appears that movie fragments are often chosen as representing scenes and actions: we could go deeper in the analysis by choosing passages where the sound design conveys specific meanings.

[3] Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *Proceedings of the 5th International Workshop on Quality of Multimedia Experience*, QoMEX, 2013

[4] `http://stefan.winkler.net/resources.html`

[5] Robin Beauchamp. *Designing Sound for Animation*. Focal Press, 2005. ISBN 0-240-80733-2

[6] Vanessa Theme Ament. *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Focal Press, 2009. ISBN 978-0-240-81125-3

Multimedia modelling

Diverse research tracks can benefit from a database of audiovisual content with eye tracking recordings: dynamic saliency models for video [7] and audio [8], movie summarization [9], interactive storytelling projects, and so on.

[7] Nicolas Riche, Matei Mancas, Dubravko Culibrk, Vladimir Crnojevic, Bernard Gosselin, and Thierry Dutoit. Dynamic saliency models and human attention: A comparative study on videos. In *Proceedings of the 11th Asian Conference on Computer Vision*, ACCV'12, 2013. DOI: 10.1007/978-3-642-37431-9_45

[8] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier. Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, 5(4):1–10, 2012

[9] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, Nov. 2013. DOI: 10.1109/TMM.2013.2267205

The sound/image relation

To underline why our dataset may be composed of diverse types of moving images related to sound, we may refer to Lev Manovich who in [10] establishes some rules of actualized cinema which he coins "soft cinema" [11]:

- "algorithmic cinema" automates the layout on the screen, the number of windows and the content;
- "macro-cinema" borrows the paradigms of computer science: windows of variable ratios and sizes;
- "multimedia cinema" considers video as one type of representation among others: 2D animation 2D, motion animation, 3D scenes, diagrams...;
- "database cinema" sources media fragments from a larger database.

[10] Lev Manovich. *The Language of New Media*. The MIT Press, 2001. ISBN 9780262133746

[11] Lev Manovich and Andreas Kratky. *Soft Cinema: navigating the database*. MIT Press, 2005. ISBN 0-262-13456-X

In one of his books [12], electroacoustic music composer and professor Michel Chion studies the relationship between sound and image. For movie analysis he coined new terms, for instance:

[12] Michel Chion. *Audio-Vision: Sound on Screen*. Columbia University Press, 1994. ISBN 0-231-07899-4

Synchresis, an acronym formed by the telescoping together of the two words synchronism and synthesis: "The spontaneous and irresistible mental fusion, completely free of any logic, that happens between a sound and a visual when these occur at exactly the same time"

Presenting results

Some researchers producing such models develop their dedicated tools to present the analysis of the database. An example is Olivier Le Meur's tool [13] [14] in Figure 1. Unfortunately, while these are sometimes offered for free upon request, most are not open source and cross-platform.

[13] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1): 251–266, 2013. DOI: 10.3758/s13428-012-0226-9

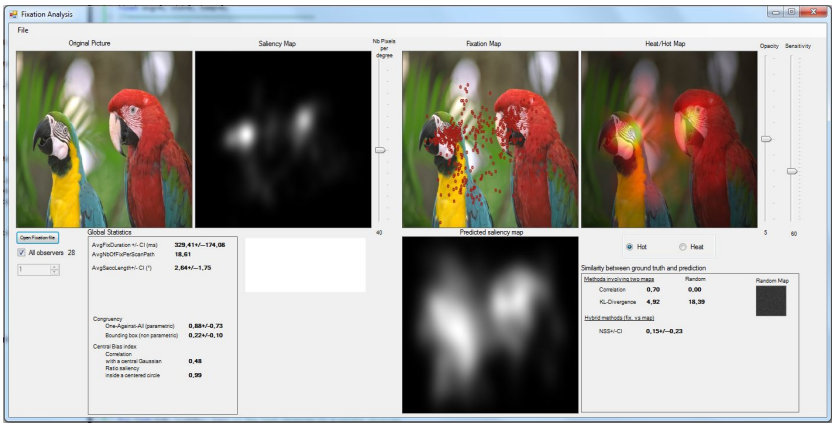

Figure 1: Saliency maps evaluation tool by Olivier Le Meur

[14] `http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html#soft`

Through a previous eNTERFACE project, some of the participants of the current project compared several multimodal annotation tools [15]. Newer interesting tools such as ChronoViz [16] appeared ever since. Such annotation tools not only allow to add metadata, but also to navigate in lengthy media content. Some of the participants are interested such video browsers, also for large collections [17].

[15] Christian Frisson, Sema Alaçam, Emirhan Coşkun, Dominik Ertl, Ceren Kayalar, Lionel Lawson, Florian Lingenfelser, and Johannes Wagner. Comediannotate: towards more usable multimedia content annotation by adapting the user interface. In *Proceedings of the eNTERFACE'10 Summer Workshop on Multimodal Interfaces*, Amsterdam, Netherlands, July 12 - August 6 2010

[16] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. Chronoviz: a system for supporting navigation of time-coded data. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 299–304. ACM, 2011. DOI: 10.1145/1979742.1979706

[17] Christian Frisson, Stéphane Dupont, Alexis Moinet, Cécile Picard-Limpens, Thierry Ravet, Xavier Siebert, and Thierry Dutoit. Videocycle: user-friendly navigation by similarity in video databases. In *Proceedings of the Multimedia Modeling Conference (MMM), Video Browser Showdown*, Huangshan, China, January 7-9 2013

Detailed technical description

Technical description

The contents of the database will be decided altogether before the workshop and composed of material released under Creative Commons licenses, particularly feature films and animation movies (for instance made with Blender) [18]. If some participants are specialists from social sciences, particularly in movie analysis, the choice of movie fragment would gain from their refinements and annotations.

For the eye tracking recordings, we will bring a Tobii Rex Developer device, but participants should feel free to bring their higher class devices.

[18] `http://wiki.creativecommons.org/Films`

Resources needed

We would need one small isolated room to perform the user tests without interruption. A pair of small powered loudspeakers for the tests would be a plus, even if we could bring it. All the remaining hardware will be provided by us.

Project management

One of the project coordinators and at least one of the proposed participants will stay for the four weeks of the workshop. This ensures at least that eye tracking experiments will be performed, that basic statistics will be extracted out of the recordings.

The part-time proposed participants will run saliency models on the database.

Based on the application of more participants, supplementary outcomes may be adapted: the database could be annotated by movie experts with any tool outputting standard formats, other methods may be investigated for the analysis of the database.

Work plan and implementation schedule

We plan the following work packages:

Gaze experiments

Preparing the test setup including an eye tracker and audiovisual content presentation tool to collect fixations from volunteers over the collection, with and without audio feedback. The related collection of audiovisual fragments will be prepared before the workshop and made available.

Multimedia modeling

Validating and improving models of audiovisual analysis (upon the participants' skills), for instance saliency.

Simple evaluation tool

Prototyping a simple application made with Qt and OpenCV to browse and compare fixations and models over the same collection. This tool might have high overlap with the gaze recording tool.

Benefits of the research

We plan to release the following deliverables:

- A database of media content and human fixations recorded through the experiments.
- The recording and comparison tools.

Team

Coordinators

Christian Frisson [19] graduated a MSc. in "Art, Science, Technology (AST)" from Institut National Polytechnique de Grenoble (INPG) and the Association for the Creation and Research on Expression Tools (ACROE), France, in 2006. Since September 2010, he is a PhD candidate with Profs Thierry Dutoit (UMONS) and Jean Vanderdonckt (UCLouvain) on the design of applications for organizating media collections (by content-based similarity) [20].

[19] http://tcts.fpms.ac.be/~frisson

[20] http://www.numediart.org/tools/mediacycle/

Nicolas Riche [21] holds an Electrical Engineering degree from the University of Mons, Engineering Faculty (since June 2010). His master thesis was performed at the University of Montreal (UdM) and dealt with automatic analysis of the articulatory parameters for the production of piano timbre. He obtained a FRIA grant for pursuing a PhD thesis about the implementation of a multimodal model of attention for real time applications [22].

[21] http://www.tcts.fpms.ac.be/homepage.php?Firstname=Nicolas&Lastname=RICHE

[22] http://tcts.fpms.ac.be/attention/

Participants

Alexis Rochette obtained his Masters in Engineering from Institut Supérieur Industriel de Bruxelles (ISIB) in 2013. He is now working there as research engineer at the Research Laboratory in the field of Arts and Sciences (LARAS) [23] on a project about interactive comic books.

[23] http://www.laras.be

Dr. Stéphane Dupont [24] received the PhD degree in Electrical Engineering at FPMs (Belgium) in 2000. He has been a visiting researcher at IDIAP (Switzerland) in 1997. Dr Dupont has also been a post-doctoral associate at the ICSI (California) in 2001-2002. There, he participated to the ETSI standardization activity on robust distributed speech recognition over wireless networks (Aurora). In 2002, he joined Multitel (Belgium), a research and innovation center, to be in charge of speech recognition research. There, he coordinated several projects, including the EU FP6 DIVINES project. He joined TCTS Lab in 2008, and is involved in the numediart programme. Dr. Dupont interests are in multimodal and speech interaction technologies, computer music, neural networks, pattern recognition and signal processing.

[24] http://www.tcts.fpms.ac.be/~dupont/

Matei Mancas [25] holds an ESIGETEL Audiovisual Systems and Networks engineering degree (Ir.), and a Orsay Univ. D.E.A. degree (MSc.) in Information Processing. He also holds a PhD in applied sciences from the FPMs on computational attention since 2007. His research deals with signal saliency and understanding. More details on Computational Attention can be found on this dedicated page [26].

[25] http://tcts.fpms.ac.be/~mancas/

[26] http://tcts.fpms.ac.be/attention/

Other researchers needed

We would welcome:

- 1 or 2 researcher(s) on social sciences studying movie making and/or sound design, to annotate the database and cue investigation tracks from the gaze recordings
- several researchers on audio and/or visual saliency modeling, or movie summarization, to analyze the database and refine their models

References

Vanessa Theme Ament. *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Focal Press, 2009. ISBN 978-0-240-81125-3.

Robin Beauchamp. *Designing Sound for Animation*. Focal Press, 2005. ISBN 0-240-80733-2.

Michel Chion. *Audio-Vision: Sound on Screen*. Columbia University Press, 1994. ISBN 0-231-07899-4.

Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier. Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, 5(4):1–10, 2012.

G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, Nov. 2013. DOI: 10.1109/TMM.2013.2267205.

Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. Chronoviz: a system for supporting navigation of time-coded data. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 299–304. ACM, 2011. DOI: 10.1145/1979742.1979706.

Christian Frisson, Sema Alaçam, Emirhan Coşkun, Dominik Ertl, Ceren Kayalar, Lionel Lawson, Florian Lingenfelser, and Johannes Wagner. Comediannotate: towards more usable multimedia content annotation by adapting the user interface. In *Proceedings of the eNTERFACE'10 Summer Workshop on Multimodal Interfaces*, Amsterdam, Netherlands, July 12 - August 6 2010.

Christian Frisson, Stéphane Dupont, Alexis Moinet, Cécile Picard-Limpens, Thierry Ravet, Xavier Siebert, and Thierry Dutoit. Videocycle: user-friendly navigation by similarity in video databases. In *Proceedings of the Multimedia Modeling Conference (MMM), Video Browser Showdown*, Huangshan, China, January 7-9 2013.

Corey Holland and Oleg Komogortsev. Eye tracking on unmodified common tablets: Challenges and solutions. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 277–280. ACM, 2012. DOI: 10.1145/2168556.2168615.

Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, 2013. DOI: 10.3758/s13428-012-0226-9.

Lev Manovich. *The Language of New Media*. The MIT Press, 2001. ISBN 9780262133746.

Lev Manovich and Andreas Kratky. *Soft Cinema: navigating the database*. MIT Press, 2005. ISBN 0-262-13456-X.

Nicolas Riche, Matei Mancas, Dubravko Culibrk, Vladimir Crnojevic, Bernard Gosselin, and Thierry Dutoit. Dynamic saliency models and human attention: A comparative study on videos. In *Proceedings of the 11th Asian Conference on Computer Vision*, ACCV'12, 2013. DOI: 10.1007/978-3-642-37431-9_45.

Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *Proceedings of the 5th International Workshop on Quality of Multimedia Experience*, QoMEX, 2013.