Auracle: how are salient cues situated in audiovisual content?

Christian Frisson, Nicolas Riche, Antoine Coutrot, Charles-Alexandre Delestage,
Stéphane Dupont, Onur Ferhat, Nathalie Guyader, Sidi Ahmed Mahmoudi, Matei Mancas,
Parag K. Mital, Alicia Prieto Echániz, François Rocca, Alexis Rochette, Willy Yvart

***Abstract*—The Auracle project aimed at investigating how sound alters the gaze behavior of people watching moving images, using low-cost opensource systems and copyleft databases to conduct experiments.**

We created a database of audiovisual content comprising: several fragments of movies with stereo audio released under Creative Commons licenses, shorts with multitrack audio shot by participants, a comic book augmented with sound (events and music). We set up a low-cost experimental system for gaze and head tracking synchronized with audiovisual stimuli using the Social Signal interpretation (SSI) framework.

We ran an eye tracking experiment on the comic book augmented with sound with 25 participants. We visualized the resulting recordings using a tool overlaying heatmaps and eye positions/saccades on the stimuli: CARPE. The heatmaps qualitatively observed don't show a significant influence of sound on eye gaze.

We proceeded with another pre-existing database of audiovisual stimuli plus related gaze tracking to perform audiovisual content analysis in order to find correlations between audiovisual and gaze features. We visualized this exploratory analysis by importing CARPE heatmap videos and audiovisual/gaze features resampled as audio files into a multitrack visualization tool originally aimed at documenting digital performances: Rekall.

We also improved a webcam-only eye tracking system, CVC Eye Tracker, by porting half of its processing stages on the GPU, a promising work to create applications relying on gaze interaction.

***Index Terms*—Eye tracking, multimodal recording, saliency, computational attention.**

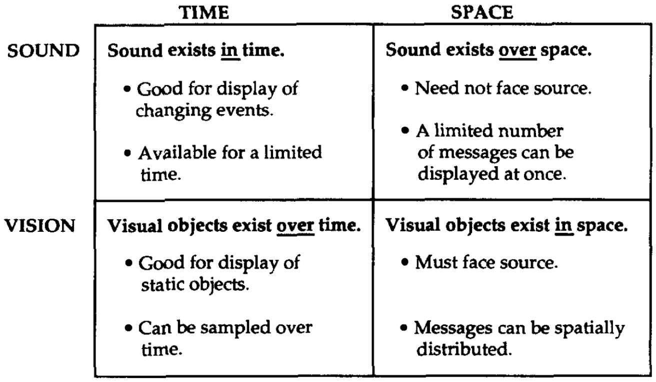
I. Introduction

Our research question is to find if gaze tracking analysis can help to understand whether or not sound influences vision when watching audiovisual content. We will first introduce facts about the sound/image relation (I-A), eye tracking (I-B) and other signals that can help to answer this question (I-C). In Section II we describe the two gaze recording setups that we put together or optimized. In Section III we illustrate the databases we used for the experiments and for the analysis. In Section IV we explain our experimental protocol for gaze recordings using an audiovisual comic book as stimulus. In Section V we report the techniques we employed to perform the analysis. In Section VI we illustrate the tools we used to produce visualization that supports our analysis.

A. The sound/image relation

The sound/image relation can be addressed by different standpoints: from the observations of the data analysts that study audio and visuals as media content or interaction channels; and from the perspective of audiovisual content creators.

Figure 1 compares the complementary modes of sound and vision across time and space as defined by Gaver for the design of human-computer interfaces [1]. It underlines the differences between audition and vision in human perception. In our case, studying the relation between sound and image implies to perform spatiotemporal analysis.



	TIME	SPACE
SOUND	**Sound exists in time.** • Good for display of changing events. • Available for a limited time.	**Sound exists over space.** • Need not face source. • A limited number of messages can be displayed at once.
VISION	**Visual objects exist over time.** • Good for display of static objects. • Can be sampled over time.	**Visual objects exist in space.** • Must face source. • Messages can be spatially distributed.

Fig. 1. Complementary modes of sound and vision by Gaver [1]

Electroacoustic music composer and professor Michel Chion has been studying the relationship between sound and image [2]. For movie analysis he coined new terms, for instance one that defines how space and time can be composed together:

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."

B. Eye tracking

Eye tracking is the measurement of the eye movements made by a viewer on a visual array, such as a real-world scene, a computer or a smartphone screen. This technique provides an unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing. Eye tracking research has been ongoing for several decades, effervescent in the last, discovered a couple of centuries ago. 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 [3].

Companies such as SensoMotoric Instruments (SMI) and Tobii provide research-grade commercial solutions that are very precise (less that 0.5° of error) and fast (samplerate in kHz) but however are very expensive (tens of thousands of Euros). Low-cost solutions with specific hardware are emerging, such as the Eye Tribe in 2003 (see Section II-B), less precise (between 0.5 and 1° of error) and fast (30 frames per second) but affordable for about hundred Euros. Low-cost solutions without specific hardware, solely based on webcams, have been developed in the research communities for a decade, such as CVC Eye-Tracker (see Section II-A) and ITU GazeTracker (evaluated by its authors [4] for usability testing versus higher-end solutions).

Besides the qualitative specification offered by dedicated hardware, another major advantage brought by expensive commercial systems is the comprehensive software tools that are provided with their product. Such tools usually cover the whole workflow for eye tracking experiments: device support, test and stimuli preparation, test recording, tests analysis with statistics. A few opensource tools providing such a fully-fledged solution exist, recently surveyed in [5], for instance OGAMA (OpenGazeAndMouseAnalyzer)[1] [6] released under a GPL license unfortunately only for Windows platforms and currently not supporting video stimuli.

C. Other signals

While the purpose of this project is to record a audiovisual database to analyze the correlation between user gaze and sound, we chose to record other signals to go deeper into the analysis. We opted for recording processed data from the Microsoft Kinect as the video signal and depth map. We have also recorded the head pose estimation and the facial "Action Units" [7] which are data for coding facial movements, for example, to know whether the user is smiling or neutral.

II. Setups

Motivated by exploring open source and low-cost solutions, we decided to investigate two eye-tracking systems. The first relying solely on a webcam is of interest to build interactive applications on mobile devices using built-in hardware (Section II-A). The second is as well low-cost but provides greater precision than the first setup for our experiments (Section II-B).

A. Setup 1: webcam-based eye tracker

Although infrared (IR) based eye-tracking techniques are used in almost all commercial eye-tracker devices and software, visible light methods pose an alternative that has the advantage of working with common hardware such as webcams. CVC Eye-Tracker [8] is an open source eye-tracking software[2] which is an improved version of Opengazer [3] and which requires no special equipment such as IR lights, IR cameras, etc.

The pipeline of the eye-tracker is shown in Fig. II-A. The application uses the images captured from the webcam to calibrate a gaze point estimator and then use this to output gaze estimations. The components of the system are:

- **Point Selection:** 8 facial feature points are selected automatically on the subject's face. A combination of Haar cascades, geometrical heuristics and a novel eye-corner detection technique is used to choose the points.
- **Point Tracking:** The points selected in the first step are tracked over the subsequent webcam frames, using a combination of optical flow and 3D head pose based estimation.
- **Calibration:** The regions containing the left and right eyes are extracted and used to calibrate a Gaussian Process (GP) estimator for the gaze point.
- **Gaze Estimation:** The left and right eye images extracted from the latest webcam frame are mapped to the screen coordinates using the GP estimator.

A key variable that has to be taken into account in real time video processing applications is the computation time, which can be so elevated in case of processing high definition (HD, Full HD, etc.) videos. Therefore, we developed a version of the eye-tracker that exploits the high computing power of graphic processing units (GPUs). These components dispose of a large number of computing units, which can be very well adapted for parallel computation. Our GPU implementation is applied on the computationally most intensive steps of the webcam-based eye tracker (Fig. II-A):

- Eyes, nose and frontal face detection
- Optical flow-based points tracking

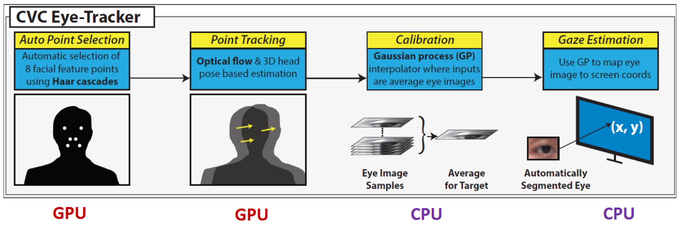

CVC Eye-Tracker
Auto Point Selection
Automatic selection of 8 facial feature points using Haar cascades
Point Tracking
Optical flow & 3D head pose based estimation
Calibration
Gaussian process (GP) interpolator where inputs are average eye images
Gaze Estimation
Use GP to map eye image to screen coords
(x, y)
Eye Image Samples
Average for Target
Automatically Segmented Eye
GPU
GPU
CPU
CPU

Fig. 2. GPU-based Eye tracker

The proposed implementation can exploit both NVIDIA and ATI graphic cards, based on CUDA[4] and OpenCL[5]. The CUDA version consists of selecting a number of GPU threads so that each thread can perform its processing on one or a group of pixels in parallel. More details about this process and the applied optimization techniques can be found in [9], [10].

Otherwise, the OpenCL implementation is based on the same process, but using a specific syntax related to OpenCL. The main advantage of OpenCL is its compatibility with both NVIDIA and ATI graphic cards, as it was proposed as a standard for GPU programming. However, CUDA, which allows to program NVIDIA cards only, offers better performances (Fig. 3) thanks to its adapted programming architecture.

[1] OGAMA: http://www.ogama.net

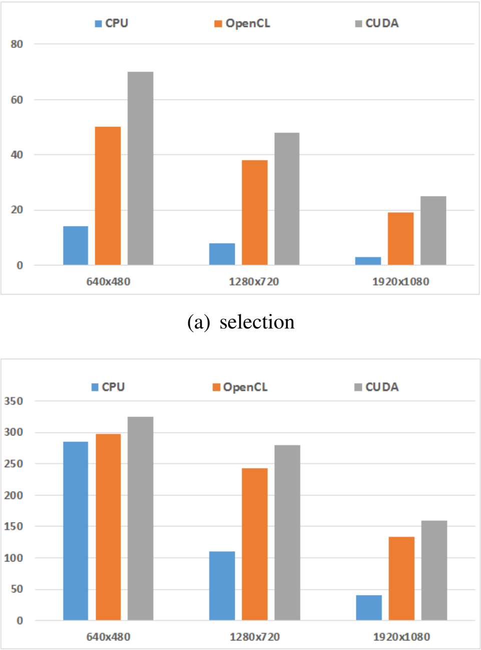
[2] CVC Eye-Tracker: http://mv.cvc.uab.es/projects/eye-tracker

[3] OpenGazer: http://www.inference.phy.cam.ac.uk/opengazer/

[4] CUDA. http://www.nvidia.com/cuda

[5] OpenCL.http://www.khronos.org/opencl

Fig. 3 compares performance between CPU, CUDA and OpenCL implementations (in terms of fps) of points (eyes, nose and frontal face) detection and optical flow-based tracking. These accelerations allowed to improve the process of webcam-based eye tracker with a factor of 3x. As result, our GPU-based method allows real time eyes tracking with high definition videos.



(a) selection

(b) tracking

Fig. 3. CUDA and OpenCL-based points selection and tracking performance

This setup does not yet provide a sufficient accuracy for performing experiments with gaze recording. In next section we describe a low-cost system that we used for our experiments.

B. Setup 2: low-cost eye tracker and depth camera

The goal of the second setup is to capture different data from a user analysis with synchronized stimuli. The recording of this data was done using the Social Signal interpretation (SSI) framework[6] developed at Augsburg University. The SSI framework gives different tools to record and analyze human behavior through sound, video, and a large range of commercial sensors [11].

In this setup, we combined a web camera, the Eye Tribe [7](a low cost eye tracking device) and the Microsoft Kinect. The utility of the webcam is to apply the CVC Eye-Tracker on the video. CVC Eye-Tracker can also be applied on the video stream from the Kinect, but with a smaller resolution. As stimuli, a video player and a comics book player are linked to the framework to obtain sensor data correctly synchronized. SSI also contains processing modules to filter and extract features from the recording signals. For example, it allows to obtain the head pose estimation and the Animated Units from the Kinect through the Kinect SDK.

The use of SSI requires writing a pipeline in XML. This pipeline contains all the structure of the system to obtain the synchronized date at the end. The XML pipeline contains 4 different parts. The first part is the sensor declaration. In this part all the sensors with the raw data extract from each sensors are declared. The second part contains filter to prepare and combine data for the visualization of the third part. The last part contains all parameters for the data storage.

Figure II-B shows the monitoring display of our pipeline recording all the sensors in sync. The next sections provide more information about the plugins implementing access to sensors and display of stimuli.

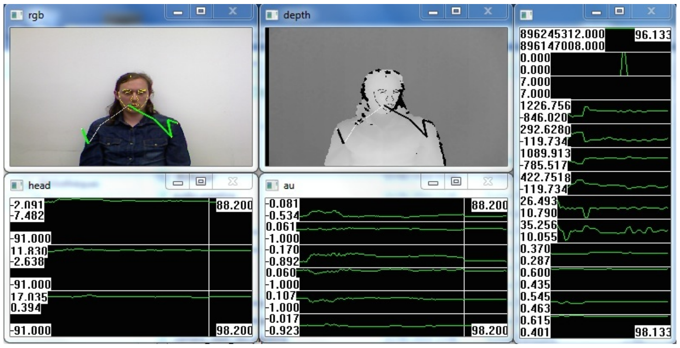


Fig. 4. Visualization of multimodal recording in SSI: Kinect RGB video (top-left), depth map video (top-middle), head pose (bottom-left), action units (bottom-middle) and Eye Tribe signals (right)

1) Webcam: The aim is to know the reactions of the user facing an audio-visual stimuli, it is normal shoot and record the user. This can easily be done using a webcam. In this case we used a 720p webcam at 30 FPS. The video was recorded using the FFMPEG Plugin from SSI

2) Kinect: To go further in the user behavior analysis, there should be an analysis of the user's face. The Kinect SSI plugin gives access to several metrics available from the Kinect SDK. Data extracted from the Kinect in this setup are the RGB Image with depth map, the head pose estimation, the user skeleton in seated mode and the facial animation unit. The Kinect sensor contains two CMOS sensors, one for the RGB image (640 x 480 pixels at 30 fps) and another for the infrared image from which the depth map is calculated, based on the deformation of an infrared projected pattern [12].

The main use of the Kinect is the user skeleton tracking. Skeletal Tracking is able to recognize users sitting. To be correctly tracked, users need to be in front of the sensor, making sure their head and upper body are visible (see Figure II-B2). The tracking quality may be affected by the image quality of these input frames (that is, darker or fuzzier frames track worse than brighter or sharp frames).

The Kinect head pose estimation method returns the Euler rotation angles in degrees for the pitch, roll and yaw as described in Figure II-B2, and the head position in meters relatively to the sensor being the origin for the coordinates.

From the face are extracted: the neutral position of the mouth, brows, eyes, and so on. The Action Units (AU) represent the difference between the actual face and the neutral face. Each AU is expressed as a weight between -1 and +1.

[6] SSI: http://www.openssi.net

[7] The Eye Tribe: http://theeyetribe.com

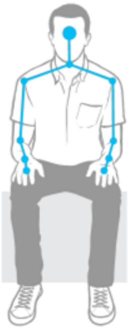


Fig. 5. User tracking in seated mode

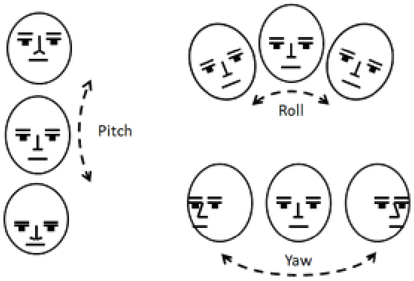


Fig. 6. Three different degrees of freedom: pitch, roll and yaw [13]

3) Eye Tribe: gaze + calibration: At the beginning of the project the data streamed by the Eye Tribe was partially recorded by SSI. We enhanced the existing Eye Tribe plugin to allow SSI to record all the gaze calibration and data provided by the Eye Tribe. The complete Eye Tribe's data description is available on their website[8]. Each participant had to perform the Eye Tribe calibration process before each stimuli type. During the calibration process SSI was recording the gaze data frame stream and the calibration data frame event. We recorded the gaze stream to have practical calibration gaze points and eventually recheck the calibration afterwards. After the launch of the calibration recording SSI pipeline we used the Eye Tribe server for the calibration which comes with Eye Tribe and consists of circular targets appearing on random grid positions on the screen. The announced Eye Tribe precision is around 0,5° to 1° which gives an average precision of 0,5cm to 1cm on the screen in practical. Practically, if the participant moves her/his head, the calibration is lost. The participant relative position to the screen and tracker is also important for a good tracking quality. The Eye Tribe and the Kinect are both using infrared illumination, these have to be properly placed to limit their interference. We used the Eye Tribe at 30 frames per second, our reference framerate for the recordings.

4) ffmpeg for movie stimuli: We wanted to get the sensors recorded in sync with the movie stimuli with two constraints: frame-accuracy and the ability to randomly sequence movies.

The SSI core plugin for video reading which uses the ffmpeg framework didn't provide the frame timing information. We modified the SSI plugin to log the frame numbers.

The ffmpeg plugin is not suitable for sequencing many video stimuli inside one pipeline. SSI provides a tool for stimuli named ModelUI that resembles a slideshow authoring tool, but not allowing frame-accurate synchronization between video stimuli (when embedded as slideshow elements) and the sensors pipeline. We had to manually generate random sequences of video stimuli by concatenating fragments (using the ffmpeg library). We experienced frame dropping through this process.

5) UDP socket client logging events from the comic book player: For the SSI synchronization of the comic book player, we simply recorded the comic book's logs through a local UDP connection. SSI provides a core plugin that reads UDP frames as events and record them in a .events xml format file. The recorded logs were: comic pages and their timestamp; sounds status (played or stopped) and type (music/effect/ambience) and sequencing option (looped or triggered once).

6) Discussion: SSI is a fully-fledged solution beyond synchronized recording providing realtime machine learning capabilities on multimodal signals.

The strengths of SSI are its availability as opensource project under a GPL license and its vast number of plugins for supporting many devices.

SSI also comes up with several drawbacks, regarding calibration and video stimuli synchronization. One calibration for both trackers (eye and kinect) was required per SSI pipeline, so we wanted to be able to launch a single pipeline per participant to facilitate the tests. From our experience with the ffmpeg plugin (see Section II-B4), we decided to drop video stimuli for the experiments (see Section IV) and choose only the interactive comic book as stimulus.

The SSI distribution (binaries and source code) that we adapter (with modified ffmpeg and Eye Tribe plugins) and used for the experiments is available on a github repository[9].

III. Database

A. State-of-the-art of audiovisual and gaze datasets

During the last decade, an exponentionally increasing number of computational visual saliency model has been proposed [14]. To be able to fairly evaluate and compare their performance ([15], [16]), over a dozen of videos datasets annotated with eye tracking data has been publicly released[10] [17]. Yet, aside from a few papers [18], [19], [20], [21], [22] authors never mention the soundtracks or explicitly remove them, making participants look at silent movies which is far from natural situations.

Here, we propose a freely available database of audiovisual stimuli and related human fixations, allowing to validate models on the relation between image and sound, with diverse genres beyond movies, documentaries and broadcasts, such as animation [23] and video games, focusing on special directing and sound design techniques [24].

[8] Eye Tribe API: http://dev.theeyetribe.com/api/

[9] SSI Auracle: http://github.com/eNTERFACE14Auracle/AuracleSSI

[10] Eye-Tracking databases repertory: http://stefan.winkler.net/resources.html

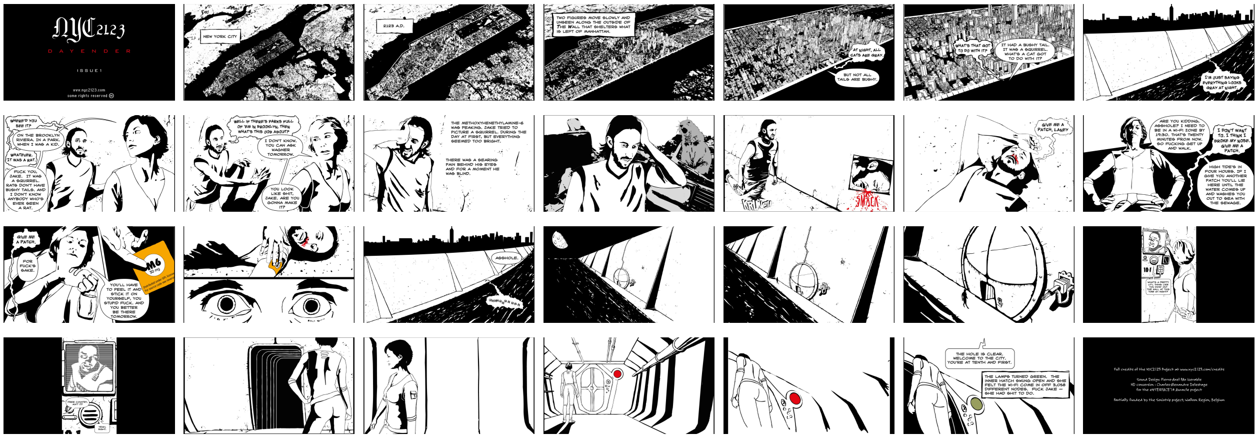

NYC2123
DAYENDER
ISSUE1
www.nyc2123.com
some rights reserved
NEW YORK CITY
2123 A.D.
TWO FIGURES MOVE SLOWLY AND UNSEEN ALONG THE OUTSIDE OF THE WALL THAT SHELTERS WHAT IS LEFT OF MANHATTAN.
AT NIGHT, ALL CATS ARE GRAY.
BUT NOT ALL TAILS ARE BUSHY.
WHAT'S THAT GOT TO DO WITH IT?
IT HAD A BUSHY TAIL. IT WAS A SQUIRREL. WHAT'S A CAT GOT TO DO WITH IT?
I'M JUST SAYING EVERYTHING LOOKS GRAY AT NIGHT.
WHERE'D YOU SEE IT?
ON THE BROOKLYN RIVIERA. IN A PARK. WHEN I WAS A KID.
WHATEVER. IT WAS A RAT.
FUCK YOU, JAKE. IT WAS A SQUIRREL. RATS DON'T HAVE BUSHY TAILS. AND I DON'T KNOW ANYBODY WHO'S EVER SEEN A RAT.
WELL IF THERE'S PARKS FULL OF 'EM IN BROOKLYN, THEN WHAT'S THIS JOB ABOUT?
I DON'T KNOW. YOU CAN ASK WAGNER TOMORROW.
YOU LOOK LIKE SHIT, JAKE. ARE YOU GONNA MAKE IT?
THE METHOXYHENETHYLAMINE-6 WAS PEAKING. JAKE TRIED TO PICTURE A SQUIRREL. DURING THE DAY AT FIRST, BUT EVERYTHING SEEMED TOO BRIGHT.
THERE WAS A SEARING PAIN BEHIND HIS EYES AND FOR A MOMENT HE WAS BLIND.
KRUNCH
SMACK
GIVE ME A PATCH, LANEY
ARE YOU KIDDING, ASSHOLE? I NEED TO BE IN A WI-FI ZONE BY 19:30. THAT'S TWENTY MINUTES FROM NOW. SO FUCKING GET UP AND WALK.
I DON'T WANT TO. I THINK I BROKE MY NOSE. GIVE ME A PATCH.
HIGH TIDE'S IN FOUR HOURS. IF I GIVE YOU ANOTHER PATCH YOU'LL LIE HERE UNTIL THE WATER COMES UP AND WASHES YOU OUT TO SEA WITH THE SEWAGE.
GIVE ME A PATCH.
FOR FUCK'S SAKE.
M6
30 mg
YOU'LL HAVE TO PEEL IT AND STICK IT ON YOURSELF, YOU STUPID FUCK. AND YOU BETTER BE THERE TOMORROW.
ASSHOLE.
UHHHHHHH HHH
10-1
WHAT'S A PRETTY LIT'L THING LIKE YOU DOIN' OUT THE WALL AT THIS TIME UH NIGHT?
FREE COUNTRY, AIN'T IT?
YEAH, RIGHT.
THE HOLE IS CLEAR. WELCOME TO THE CITY. YOU'RE AT TENTH AND FIRST.
THE LAMPS TURNED GREEN. THE INNER HATCH SWUNG OPEN AND SHE FELT THE WI-FI COME IN OFF 3,058 DIFFERENT NODES. FUCK JAKE — SHE HAD SHIT TO DO.
Full credits of the NYC2123 Project at www.nyc2123.com/credits
Sound Design: Pierre-Axel PAx Izerable
HD conversion : Charles-Alexandre Delestage
for the eNTERFACE'14 Auracle project
Partially funded by the Sonixtrip project, Walloon Region, Belgium

Fig. 7. Frames of the NYC2123 comic book reworked with less text, as audiovisual stimulus for the experiments

B. The Auracle database

1) Movies with multitrack audio: To provide material for the experiments following our initial plans to include movie stimuli, 8 excerpts from Creative Commons movies were selected for the added values sound offered against audio, with an access to the source files in order to modify the sound channels. We also wrote and shot 5 original shorts with a cinema-like aspect. Each of these had separate audio channels for music, ambience and effects, so the sound mixing could be changed during the experiments using a control surface.

These movies could unfortunately not be used as stimuli for experiments due to our issues with implementing a suitable SSI pipeline with frame-accurate synchronization between video stimuli and sensor recording (see Section II-B4).

2) A comic book augmented with sound: We used a digital comic book with sound effects and music as a stimuli. We wanted that all the material (comic book, sound effects, music) would be released under a Creative Commons license. We adapted the source files of the NYC2123 comic book[11] to make all pages match a single orientation and resolution. Pierre-Axel Izerable produced the sound design and provided the audio files and a description of their sequencing. We assembled the visuals and the sounds together in a prototype with which sounds can be looped or played once, in transition between two frames. Participants can read the comic book at their own pace, the prototype records the slide transitions as well as if an audio file is played, stopped, or looped; the type of audio content of the file (ambient, sound effect or music). This information is also sent to the SSI recorder through an UDP connection so it can be synchronized with the sensors data. After some testing we decided to remove some large framed text from the original comic because the participants were taking a huge proportion of the experiment duration to read those and the reading path and time were not the gaze features we wanted to extract. Fig. 7 illustrates these reworked frames containing less text. An HTML5 version of the NYC2123 comic book augmented with sound is available online[12].

IV. Experiments

A. Participants

25 participants agreed to take an experiment consisting in watching the comic book while their gaze was being monitored. Participants were split into two groups: one group experienced the comic book augmented with sound, the other group read the comic book without soundtrack. Figure IV-A shows the apparatus used for the experiments.

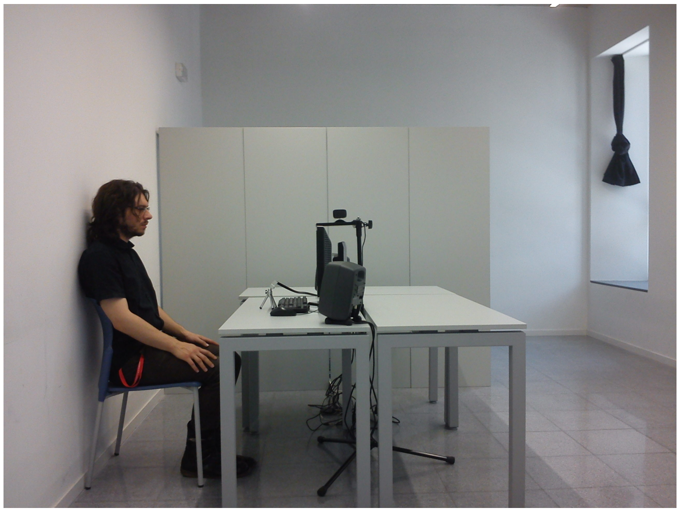


Fig. 8. Apparatus used for the experiments

Participants didn't receive financial reward but were offered Belgian chocolate.

We used the opensource survey management application LimeSurvey[13] to implement and collect pre- and post-experiment questionnaires. The pre-experiment questionnaire allowed to collect qualitative data: sex, spoken language(s). Among the 25 participants, 6 were females and 19 males. 6 reported English as their native language. 12 reported to suffer from visual impairment: 11 included left/right eye myopia/presbyopia or strabismus; and 1 color blindness. 3 reported to suffer from hearing impairment, including slight loss in bass/medium/treble frequency range or tinnitus.

[11] NYC2123 comic book: http://nyc2123.com

[12] Auracle NYC2123 Issue 1 augmented with sound: http://github.com/eNTERFACE14Auracle/AuracleNYC2123Issue1

[13] LimeSurvey: http://www.limesurvey.org

B. Collected data

For each participant, data was collected in files containing:

- events from the comic book player with timestamps: next comic book frame, sound start/loop/end;
- streams from the trackers: Kinect RGB/depthmap videos and head pose and action units, EyeTribe eye fixation status and raw/average eye positions;
- calibration status from the Eye Tribe including an estimation of its quality on a [0;5] scale.

C. Preliminary analysis on reading times

In this Section, we perform analysis on temporal aspects: reading times that can be grouped by participant, by comic book frame, and by condition (with or without sound).

1) User-defined page switching pace complicates experimental analysis: Participants could read the comic book at their own pace, they could decide when to switch to the next page manually by pressing the right arrow key on a keyboard.

Table I visualizes the variations in reading time between participants and on each comic book frame. For each participant, a replay video was generated with a fixed framerate, along the time spent reading each comic book frame. Each of these replay videos was summarized into a slit-scan (by concatenating horizontally a 1-pixel-wide central vertical column extracted in each frame).

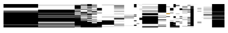
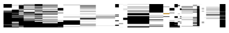
id	av	comicstripscan
3	0	
17	0	
24	1	
4	1	
2	0	
21	0	
10	1	
1	1	
7	0	
19	0	
0	1	
8	1	
9	0	
14	1	
13	0	
11	0	
20	1	
23	1	
22	0	
12	1	
18	1	
6	1	
15	1	
16	0	
5	0	

TABLE I
COMICSTRIPSCANS ORDERED PER DURATION AGAINST PARTICIPANT ID AND AV CONDITION (WITH OR WITHOUT SOUND)

This illustrates how our stimuli have a variable duration depending on the browsing pace desired by each participant, what makes the inter-subject analysis more difficult.

Figure 9 visualizes times per frame grouped by user. Besides users 5, 6, 15 and 16 (at the bottom of Table I), most users read most frames in between 2 and 10 seconds, except for some frames considered as outliers. Analyzing the effect of frame content on reading times may help to further understand this.

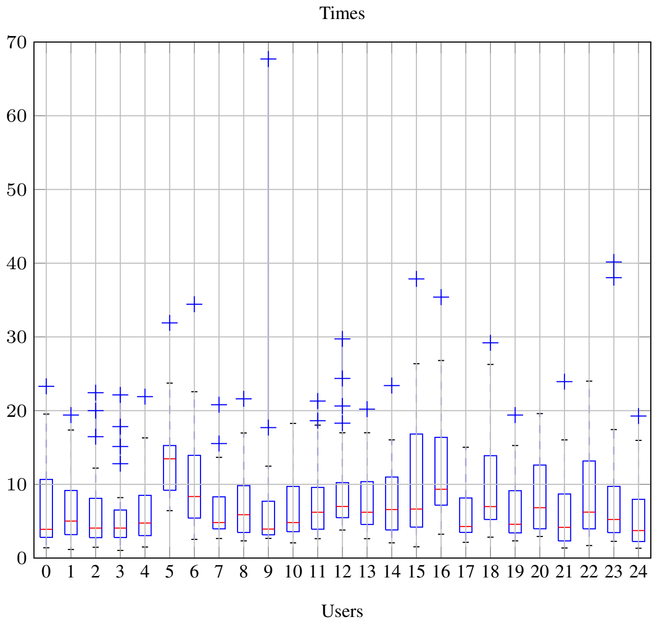


Fig. 9. Times (s) per frame grouped by user

2) Influence of frame content on reading times: Figure 10 visualizes times per user grouped by frame. Reading times are clearly greater for frames containing much text (7-9, 13-14, 26). Further analysis may be undertaken to correlate text length and complexity with reading times.

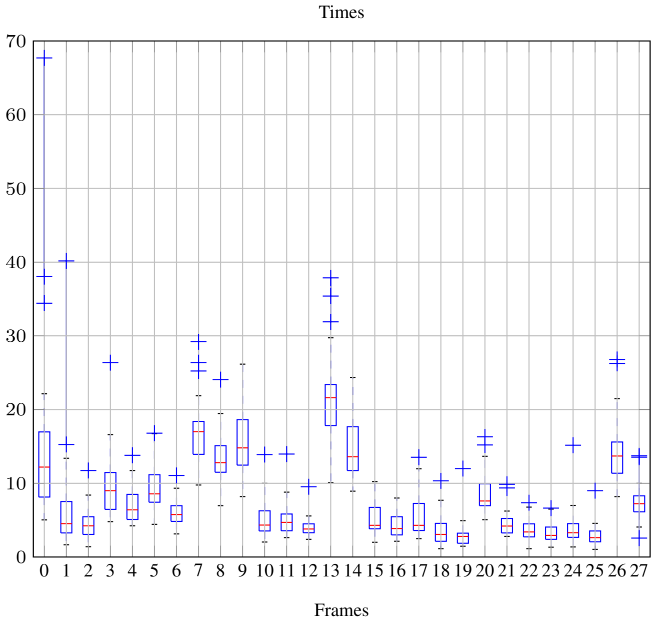


Fig. 10. Times (s) per user grouped by frame

3) Influence of the audio condition (with/without soundtrack) on reading times: The Shapiro-Wilk test statistic (W) and its p-value (p) can be computed to to test the null hypothesis whether reading times (for all frames and participants) grouped by both conditions (with and without sound) are taken from a normal distribution [25]. Times "with sound" (W=0.82, p=1.15e-19) and "without sound" (W=0.77, p=1.55e-21) are not from a normal distribution (the null hypothesis is rejected).

We want to compare two unpaired groups (different amount of participants and varying times per frame) whose distribution normality is not assumed, therefore we choose to compute the Mann-Whitney test statistic (u) and its p-value (p) [26]. With u=6.11e5 and p=0.49, we can assume that there is no signifiant difference in reading times per frame per participant between both groups (with/without audio), what Figure 11 illustrates.

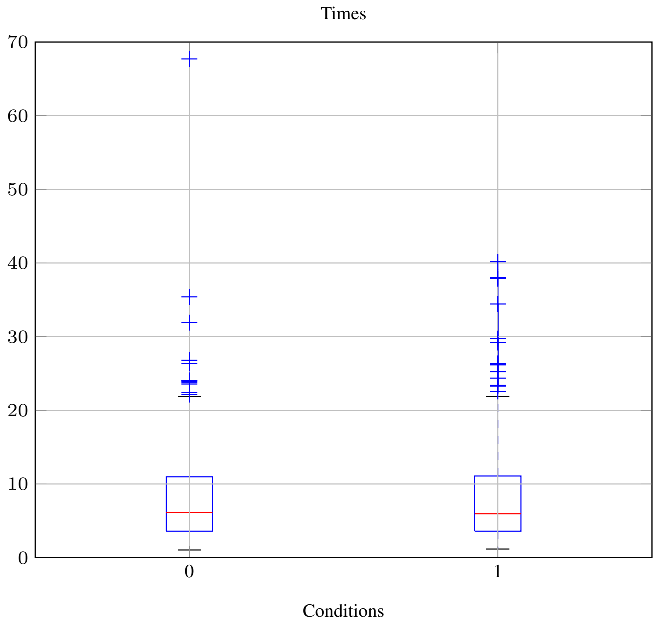


Fig. 11. Times (s) per user per frame grouped by condition (0: without sound, 1: with sound)

This experiment could only be undertaken during the last week of the workshop, after facing technical difficulties on synchronized recording and thus deciding to simplify the test protocol discarding video stimuli. We used a pre-existing database of video stimuli and related gaze recordings by Coutrot el al. [19] to test and choose techniques for analysis (Section V) and visualization (Section VI) .

V. Analysis

In this section, we report how audiovisual content and gaze recordings can be analyzed; and how the analyses of both can be correlated.

A. Audiovisual content analysis

Audio content can be described by a collection of commonly employed musical information retrieval features: the first 4 central moments of the real portion of a fast Fourier transform, spectral flux, entropy, first 13 MFCCs, and their flux. These are computed using the default parameters from MIRtoolbox [27]. Alternative methods could be employed which attempt to describe the content by its latent feature dimensions [28].

Regarding visual features, we compute optical flow, shown to correlate significantly with the entropy of eye-movements from multiple participants during dynamic scene viewing [29], [30]. This feature describes a vector of motion for each pixel for every frame. The resulting vectors are binned in a 360 degree histogram for each frame, creating a 360-element feature vector.

B. Gaze recordings analysis

We decided to describe gaze by the dispersion of the eye-movements, following the tracks of Coutrot et al [19]. Dispersion is defined as the mean of the Euclidean distance between the eye positions of different observers on a given frame.

C. Correlations between audiovisual/gaze features

Correlating two multivariate features can be done by a canonical correlation analysis. This method attempts to describe either multivariate feature as a linear combination of the other. By reducing the data to single linear combination, we can produce a single metric describing how well one domain is described by the other. The resulting feature can be used in a program such as Rekall (see Section VI-B) for alternative methods of linear editing, producing meaningful thumbnails, or simply providing the editor another cue for understanding their media. When dealing with larger collections, this feature could possibly be used for align/sync audiovisual content.

VI. Visualization supporting analysis

Techniques for visualizing eye tracking have been elaborated since the 1950, as surveyed in the state-of-the-art report by Blascheck et al.[31]. Some of these visualization techniques are integrated into tools that allow exploratory analysis and interactive browsing, such as the recent and comprehensive tool ISeeCube by Kurzhals et al.[32]. Unfortunately, a majority of these tools are neither open source nor cross-platform. We opted for combining two opensource solutions: CARPE (Section VI-A) that computes heatmaps and scanpaths, and Rekall (Section VI-B) that allows to browse audio and video content through a multitrack timeline.

A. Gaze heatmaps and scanpaths with CARPE

CARPE [29] produces heatmaps overlaid on video and imports eye-tracking data stored in text files. Each eye-movement is convolved with a 2-degree isotropic Gaussian distribution. The resulting distribution is normalized and converted to a jet-colormap. The CARPE source code is available on a github repository[14]. We use CARPE for producing example visualizations of our data. A heatmap video rendered with CARPE with the gaze data recorded through the experiments on the interactive version of the NYC2123 comic book is available online[15].

[14] CARPE: https://github.com/pkmital/NSCARPE

[15] CARPE replay videos of Auracle NYC2123 Issue 1: https://archive.org/details/AuracleNYC2123CARPE

B. Multitrack audiovisual/gaze visualization with Rekall

Tools that allow exploratory browsing of audiovisual content are multimodal annotation tools (surveyed by some of the participants of the current project through a previous eNTERFACE project [33]) and non-linear audio/video editing tools and sequencers. Few of these tools are cross-platform, opensource and easy to adapt to our needs. Rekall [16] is a new opensource tool for browsing and annotating audiovisual content, designed for the documentation of live artistic performances [34]. It is released under an opensource license (CeCILL, similar to GPL). Until the end of the eNTERFACE'14 workshop, its first version was developed as a desktop application solely relying on the Qt framework[17], pre-computing and visualizing simple thumbnails such as audio waveforms and video keyframes. Rekall was thus a good candidate for our requirements (opensource, cross-platform, easy to adapt).

Figure 12 illustrates how Rekall can be adapted to visualize altogether a stimulus video with a heatmap of gaze fixations overlaid using CARPE and multitrack thumbnails of features (dispersion and audiovisual correlation) and video (keyframes). Such a prototype has been obtained throughout the workshop with a limitation: some of the files recorded through the experiments are converted to files supported by Rekall, rather than having implemented the computation of their visualization directly into Rekall. These are the heatmap videos pre-computed using CARPE, importable as video content by Rekall; and the audiovisual/gaze features resampled at a minimal standard audio sampling rate (22050 Hz) to be directly importable as audio content from which Rekall computes waveforms.

Next steps would include to adapt Rekall to support various and modular thumbnail techniques through a plugin architecture, so that a single standalone application can be used without external conversions.

VII. Conclusion and Future Work

Throughout this project we have learned by trial and error the current state of the art on: low-cost eye tracking, synchronized multimodal recording, opensource eye tracking visualization and analysis.

We applied this experimental pipeline to an experiment with 25 participants whose gaze was tracked while watching an interactive comic book augmented with sound. The collected data is not relevant for further analysis, therefore we decided not to distribute it. However the experiment can be replicated on different datasets using our system.

Further work is required to allow the analysis of the correlation between the direction of the head and eye tracking.

To better understand the zones of interest in each comic book frame, clues could be mined from visual analysis, such as *objectness* to track protagonists and objects in the scene [35], and text detection and recognition [36].

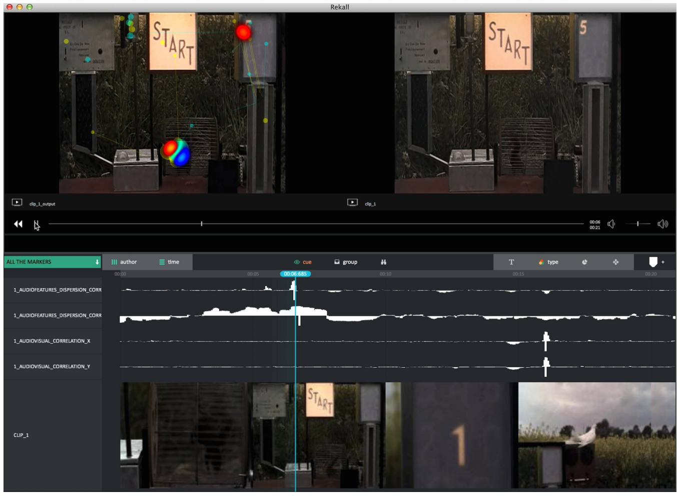

Rekall
clip_1_output
clip_1
00:06
00:21
ALL THE MARKERS
author
time
cue
group
type
00:00
00:05
00:06:685
00:10
00:15
00:20
1_AUDIOFEATURES_DISPERSION_CORR
1_AUDIOFEATURES_DISPERSION_CORR
1_AUDIOVISUAL_CORRELATION_X
1_AUDIOVISUAL_CORRELATION_Y
CLIP_1

Fig. 12. Visualization in Rekall of a stimulus video (top-right) with a heatmap of gaze fixations overlaid using CARPE (top-left) and multitrack thumbnails (below) of features (dispersion and audiovisual correlation) and video (keyframes)

Acknowledgment

Christian Frisson, Alexis Rochette and Stéphane Dupont have been partially funded by a grant from the Walloon Region of Belgium, GREENTIC program, entitled SonixTrip (number 1317970) (http://www.sonixtrip.be) for the preparation of this project and their participation to eNTERFACE'14.

Christian Frisson received a grant from the International Speech Communication Association (ISCA) to partially cover his expenses during his participation to the eNTERFACE'14 workshop (http://www.isca-speech.org).

Onur Ferhat's work was supported in part by the Spanish Gov. grants MICINN TIN2009-10435 and Consolider 2010 MIPRCV, and the UAB grants.

Sidi Ahmed Mahmoudi has been partially funded by the computer science department of the faculty of Engineering of UMONS. Sidi Ahmed Mahmoudi would like to thank the PPP Slowdio project for supporting his research activities.

François Rocca is supported by the Integrated Project LinkedTV (www.linkedtv.eu) funded by the European Commission through the 7th Framework Programme (FP7-287911).

The authors would like to thank the local organizers of eNTEFACE'14 for their kindness, their local assistance regarding logistics, their nice schedule of social events, their successful call of ISCA grants to support participating students.

The NYC2123 comic book (http://nyc2123.com) and its Adobe Illustrator source files created by Chad Allen and Paco Allen are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License.

The authors would like to thank Pierre-Axel Izerable for having produced the sound design for the comic book and accepted to release it under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License.

The authors would like to thank Johannes Wagner for having granted them access to their bleeding-edge repository for the SSI framework [11].

[16] Rekall: http://www.rekall.fr – http://github.com/Rekall/Rekall

[17] Qt: http://qt-project.org

The authors would like to thank Jerome Urbain for his introduction and help on SSI pipeline creation.

The authors would like to thank Marc Thunissen from Océ Software Laboratories for the HTML5/CSS/JS code that allowed us to replicate the NYC2123 Issue 1 comic book augmented with sound.

Christian Frisson would like to thank the creators of Rekall for the insightful discussions they shared about their tool: Clarisse Bardiot, Guillaume Jacquemin, Guillaume Marais and Thierry Coduys.

Christian Frisson would like to thank Serheyev Pavlo for having trustfully shipped his for sale EyeTribe eye tracker before he could receive the payment, while official sources were out of stock. This helped us getting started.

References

[1] W. W. Gaver, "The sonic finder: An interface that uses auditory icons," *Human-Computer Interaction*, vol. 4, pp. 67–94, 1989.

[2] M. Chion, *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

[3] C. Holland and O. Komogortsev, "Eye tracking on unmodified common tablets: Challenges and solutions," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '12. ACM, 2012, pp. 277–280.

[4] S. A. Johansen, J. San Agustin, H. Skovsgaard, J. P. Hansen, and M. Tall, "Low cost vs. high-end eye tracking for usability testing," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 1177–1182.

[5] V. Krassanakis, V. Filippakopoulou, and B. Nakos, "Eyemmv toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification," *Journal of Eye Movement Research*, vol. 7, no. 1, pp. 1–10, 2014.

[6] A. Vokhler, V. Nordmeier, L. Kuchinke, and A. M. Jacobs, "Ogama - OpenGazeAndMouseAnalyzer: Open source software designed to analyze eye and mouse movements in slideshow study designs," *Behavior Research Methods*, vol. 40, no. 4, pp. 1150–1162, 2008.

[7] P. Ekman, W. Friesen, and J. Hager, *The Facial Action Coding System*. The MIT Press, 2002.

[8] O. Ferhat, F. Vilariño, and F. J. Sanchez, "A cheap portable eye-tracker solution for common setups," *Journal of Eye Movement Research*, vol. 7, no. 3, pp. 1 – 10, 2014.

[9] S. A. Mahmoudi, M. Kierzynka, P. Manneback, and K. Kurowski, "Real-time motion tracking using optical flow on multiple gpus," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 62, pp. 139–150, 2014.

[10] S. A. Mahmoudi, M. Kierzynka, and P. Manneback, "Real-time gpu-based motion detection and tracking using full hd videos," in *Intelligent Technologies for Interactive Entertainment*, 2013, pp. 12–21.

[11] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 831–834.

[12] H. Gonzalez-Jorge, B. Riveiro, E. Vazquez-Fernandez, J. Martínez-Sánchez, and P. Arias, "Metrological evaluation of microsoft kinect and asus xtion sensors," *Measurement*, vol. 46, no. 6, pp. 1800–1806, 2013.

[13] "Face tracking. microsoft developper network." [Online]. Available: http://msdn.microsoft.com/en-us/library/jj130970.aspx

[14] A. Borji and L. Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[15] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and Human Fixations: State-of-the-art and Study of Comparison Metrics," in *Proceedings of the 14th International Conference on Computer Vision (ICCV 2013)*, Sydney, Australia, 2013, pp. 1–8.

[16] ——, "A study of parameters affecting visual saliency assessment," in *Proceedings of the 6th International Symposium on Attention in Cognitive Systems (ISACS'13)*, Beijing, China, 2013.

[17] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience*, ser. QoMEX, 2013.

[18] C. Quigley, S. Onat, S. Harding, M. Cooke, and P. König, "Audio-visual integration during overt visual attention," *Journal of Eye Movement Research*, vol. 1, no. 2, pp. 1–17, 2008.

[19] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[20] M. L. H. Võ, T. J. Smith, P. K. Mital, and J. M. Henderson, "Do the eyes really have it? Dynamic allocation of attention when viewing moving faces," *Journal of Vision*, vol. 12, no. 13, pp. 1–14, 2012.

[21] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[22] ——, "An Audiovisual Attention Model for Natural Conversation Scenes," in *IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014.

[23] R. Beauchamp, *Designing Sound for Animation*. Focal Press, 2005.

[24] V. T. Ament, *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Focal Press, 2009.

[25] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[26] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, p. 5060, 1947.

[27] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," *International Conference on Digital Audio Effects*, pp. 237–244, 2007.

[28] P. Mital and M. Grierson, "Mining Unlabeled Electronic Music Databases through 3D Interactive Visualization of Latent Component Relationships," in *NIME 2013: New Interfaces for Musical Expression*, Seoul, Korea, 2013.

[29] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, Oct. 2010.

[30] T. J. Smith and P. K. Mital, "Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes," *Journal of Vision*, vol. 13, pp. 1–24, 2013.

[31] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-Art of Visualization for Eye Tracking Data," in *State of The Art Report (STAR) from the Eurographics Conference on Visualization (EuroVis)*, R. Borgo, R. Maciejewski, and I. Viola, Eds. The Eurographics Association, 2014.

[32] K. Kurzhals, F. Heimerl, and D. Weiskopf, "Iseecube: Visual analysis of gaze data for video," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14. New York, NY, USA: ACM, 2014, pp. 351–358.

[33] C. Frisson, S. Alaçam, E. Coşkun, D. Ertl, C. Kayalar, L. Lawson, F. Lingenfelser, and J. 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.

[34] C. Bardiot, T. Coduys, G. Jacquemin, and G. Marais, "Rekall: un environnement open source pour documenter, analyser les processus de creation et faciliter la reprise des uvres sceniques," in *Actes des Journées d'Informatique Musicale*, 2014.

[35] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014.

[36] L. Neumann and M. J, "Real-time scene text localization and recognition," in *IEEE CVPR*, 2012.

Christian Frisson 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. In 2015, he obtained his PhD with Profs Thierry Dutoit (UMONS) and Jean Vanderdonckt (UCL) on designing interaction for organizating media collections (by content-based similarity). He has been a fulltime contributor to the numediart Institute since 2008.
http://christian.frisson.re

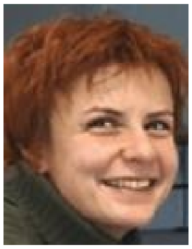
Nicolas Riche holds an Electrical Engineering degree from the University of Mons, Engineering Faculty (since June 2010). His master thesis was performed at the University of Montreal (UdM) and dealt with automatic analysis of the articulatory parameters for the production of piano timbre. He obtained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

Antoine Coutrot holds an Engineering degree and a PhD in Cognitive Science from Grenoble University (France). During his PhD, he studied the influence of sound on the visual exploration of dynamic scenes. He currently holds a Research Associate position at CoMPLEX, University College London (UK), where he develops statistical model to understand how high and low level features guide our gaze.
http://www.gipsa-lab.fr/~antoine.coutrot

Charles-Alexandre Delestage is an Msc student in Audiovisual Communication's Management in Valenciennes. He starting a doctorate degree fall 2014 within DeVisu (France) and TCTS (Belgium) research laboratories. His research is oriented on the automation in the audiovisual processes of production and the impact on the audiences.
http://www.univ-valenciennes.fr/DEVISU/membres/delestage_charles_alexandre

Stéphane Dupont received the PhD degree in EE at the Faculty of Engineering of Mons (Belgium) in 2000. He was post-doctoral associate at ICSI (California) in 2001-2002 where he participated to the ETSI standardisation activity on robust speech recognition. In 2002, he joined Multitel (Belgium) to coordinate the speech recognition group and several industrial and EU-funded projects. In 2008, he joined UMONS (Belgium). Dr. Dupont interests are in speech, music, audio and multimedia processing, machine learning, and multimodal human-computer interaction. He holds 3 international patents and has authored/co-authored over 100 papers in these areas.
http://www.tcts.fpms.ac.be/~dupont

Onur Ferhat holds a BSc degree in Computer Engineering from Boğaziçi University, Turkey and an MSc degree in Computer Vision from Universitat Autònoma de Barcelona (UAB). He is currently a postgraduate student in Computer Vision Center, Barcelona and an assistant teacher in UAB. His research interests include computer vision, eye-tracking and human-computer interaction.
http://onurferhat.com



Nathalie Guyader obtained a PhD degree in Cognitive Sciences at the University Joseph Fourier of Grenoble (France) in 2004 under the supervision of J. Hérault and C. Marendaz on a biologically inspired model of human visual perception to categorize natural scene images. Since 2006, she has been an associate professor at the University Joseph Fourier of Grenoble and at the Grenoble Image Signal and Automatism laboratory (GIPSA-lab). Her research mainly concerns human visual attention through two approaches : eye-tracking experiment analysis and computational modeling.
http://www.gipsa-lab.grenoble-inp.fr/page_pro.php?vid=98

Sidi Ahmed Mahmoudi received the graduate engineering degree in computer science from the University of Tlemcen, Algeria, the masters degree in multimedia processing from the Faculty of Engineering in Tours, France, and the PhD degree in engineering science from the University of Mons, Belgium, in 2006, 2008, and 2013, respectively. Currently, he is a postdoc researcher at the University of Mons, Belgium. His research interests are focused on real time audio and video processing for watching slow motions, efficient exploitation of parallel (GPU) and heterogeneous (multi-CPU/multi-GPU) architectures. He also participated in national (ARC-OLIMP, Numdiart, Slowdio PPP) projects and European actions (COST IC 805).
http://www.ig.fpms.ac.be/fr/users/mahmoudisi

Matei Mancas 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. Part of this research is done for artisitic purposes within the numediart institute. Matei is now the SmartSpaces Group leader @ the Numediart Research Institute (UMONS Institute for Creative Technologies).
http://tcts.fpms.ac.be/~mancas

Parag K. Mital B.Sc, M.Sc., Ph.D, is a computational media artist investigating how film together with eye-movements, EEG, and fMRI can help to explore how people attend to and represent audiovisual scenes. His arts practice builds this understanding into audiovisual scene synthesis, computational mashups, and expressive control of audiovisual content. He is currently working on audiovisual decoding with fMRI at Dartmouth College, Bregman Music and Audio Research Studio.
http://pkmital.com

Alicia Prieto Echániz graduated in Advertising and Public Relations and has a master's degree in Corporate Communications from the University of the Basque Country. Specialized on photography, graphic design and online marketing, has taken part in three short film contests and has juggled her studies with voluntary work in brain injury daycare centers and as a instructor in urban summer camps. She's currently working for the Cultural Association Tarasu as the communication's manager, which dedicates to the spread of the Japanese culture.
http://es.linkedin.com/pub/alicia-prieto-echniz/33/955/3/en

François Rocca holds an Electrical Engineering degree from the Faculty of Engineering of Mons (UMONS) since June 2011. He did his master's thesis in the field of emotional speech analysis, and more especially on laughter frequencies estimation. He is currently pursuing a PhD thesis on Real-time 2d/3D head pose estimation, face tracking and analysis by markerless motion tracking for expression recognition.
http://tcts.fpms.ac.be/∼rocca

Alexis Rochette is currently researcher on the SonixTrip project for the art and science laboratory (Laras) of Institut de Recherche de l'Institut Supérieur Industriel de Bruxelles (IrIsib). He holds an Industrial Electronic Engineering degree from Institut Supérieur Industriel de Bruxelles (Isib) since 2013. He did a Final internship for the Master Degree and Master thesis at Klavis Technologies S.A from Brussels where he developed of a CopperLan application on a embedded platform. He also holds a Bachelor Degree in Applied Electronics since 2011 for which he did a Final internship at the Royal Observatory of Belgium.

Willy Yvart graduated a Master Degree in Multimedia, Audiovisual, Information and Communication Sciences from DREAM departement of the University of Valenciennes (France) in 2011. Since 2013, he is a PhD candidate under the joint supervision of Thierry Dutoit (UMONS, Belgium) and Sylvie Leleu-Merviel (UVHC, France) on the study of semantics metadata in massive music library in order to improve indexing and searching techniques.
http://www.univ-valenciennes.fr/DEVISU/membres/yvart_willy