AUDIO-VISUAL ATTENTION: EYE-TRACKING DATASET AND ANALYSIS TOOLBOX

Pierre Marighetto[1] *Antoine Coutrot*[2] *Nicolas Riche*[1] *Nathalie Guyader*[3]
Matei Mancas[2] *Bernard Gosselin*[1] *Robert Laganiere*[4]

[1]University of Mons, Belgium
[2]University College London, UK
[3]GIPSA-lab, University of Grenoble, France
[4]University of Ottawa, Canada

ABSTRACT

Although many visual attention models have been proposed, very few saliency models investigated the impact of audio information. To develop audio-visual attention models, researchers need to have a ground truth of eye movements recorded while exploring complex natural scenes in different audio conditions. They also need tools to compare eye movements and gaze patterns between these different audio conditions. This paper describes a toolbox that answer these needs by proposing a new eye-tracking dataset and its associated analysis ToolBox that contains common metrics to analysis eye movements. Our eye-tracking dataset contains the eye positions gathered during four eye-tracking experiments. A total of 176 observers were recorded while exploring 148 videos (mean duration = 22 s) split between different audio conditions (with or without sound) and visual categories (moving objects, landscapes and faces). Our ToolBox allows to visualize the temporal evolution of different metrics computed from the recorded eye positions. Both dataset and ToolBox are freely available to help design and assess visual saliency models for audiovisual dynamic stimuli.

Index Terms— Visual Attention, Audio-visual Attention, Saliency, Videos, Multi-modal, Eye-tracking Database, ToolBox

1. INTRODUCTION

Human visual attention is not only driven by visual stimuli but also by other modalities, such as auditory stimuli. Hearing and sight constantly interact to perceive the surrounding world. Audiovisual illusions are certainly the most popular audiovisual interactions. For instance, the McGurk effect appears when mismatched acoustic and visual stimuli are simultaneously presented. The result is a perceptual shift: auditory /ba/ and visual /ga/ are audiovisually perceived as /da/ [1]. The Superior Colliculus (SC) integrates information coming from different sensory areas [2]. Visual information is processed on the superficial layers and the auditory information on the deeper layers [3]. Once in the same coordinate system, multi-sensory information is fused in order to guide the eyes onto the audiovisually most *salient* areas of our visual field.

Despite the ubiquitousness of sound in real life, gaze-based studies rarely take this information into account. Authors often record eye movements of observers exploring silent dynamic scenes, which is far from natural situations. Recent studies [4, 5, 6] confirmed the impact of soundtrack on gaze while watching videos. Thus, not considering sound in attention modeling induces some serious limitations, which will only be overcome with a new generation of multimodal models.

There is no easy way to introduce auditory information into dynamic visual attention models [7]. A few multimodal saliency models have been proposed, but for specific topics such as video summarization [8] or conversation scenes [9, 10]. In this context, we aim at building an open source general framework containing a visual vs audio-visual eye-tracking and video database. We also provide a ToolBox allowing to easily compute and visualize metrics computed from the recorded eye positions. The interests of this work are two-fold. First, it enables a frame-by-frame quantification of the impact of sound while exploring videos. Second, it provides a large and diverse ground-truth for audiovisual saliency models assessment.

This paper is organized as follows: Section 2 gives details about the database architecture, Section 3 details the metrics available in the associated analysis ToolBox and Section 4 provides an example of the possibilities offered by the ToolBox on the database. Finally, Section 5 concludes the paper.

2. EYE-TRACKING AND VIDEO DATABASE

The database has been built through four different experiments. The purpose is to obtain a consistent, homogeneous and reliable dataset. The design of the first experiment is ex-

haustively presented. For the two following experiments, only diverging points are reported. The fourth one is an original experiment of this paper.

2.1. First Experiment

This dataset has been used in [4].

- Participants: 40 participants: 26 men and 14 women, from 20 to 29 years (M = 25.3; SD = 2.7).

- Stimuli: 50 varied videos extracted from professional movies (action movies, drama, documentary films, dialogues), with their original monophonic soundtracks (48 kHz). When the soundtrack contained speech, it was always in French. Each video sequence has a resolution of 720 x 576 pixels (30 x 24 degrees of visual angle) and a frame rate of 25 frames per second. They last from 0.9 s to 35 s (M = 8.7 s; SD = 7.2 s).

- Protocol: Participants had to look freely at the 50 videos. In order to avoid any order effect, videos were randomly displayed. Twenty participants saw the first half of videos in the visual condition (i.e. without any sound) and the other half in the audio-visual condition (i.e. with their original soundtracks), with a small break in between. Stimulus conditions (Visual and Audio-Visual) were counterbalanced between participants. Each experiment was preceded by a calibration procedure, during which participants focused their gaze on 9 separate targets in a 3 x 3 grid that occupied the entire display. A drift correction was carried out between each video, and a new calibration was done at the middle of the experiment and if the drift error was above 0.5 degree.

- Apparatus: Eye movements were recorded using an eyetracker (Eyelink 1000, SR Research) with a sampling rate of 1000 Hz in binocular pupil - corneal reflect tracking mode. Participants were seated 57 cm away from a 21 inch CRT monitor with a spatial resolution of 1024 x 768 pixels and a refresh rate of 75 Hz. The audio signal was presented via headphones (HD280 Pro, 64 Ohm, Sennheiser). Participants wore headphones during the whole experiment, even when the stimuli were presented without soundtrack.

2.2. Second Experiment

This dataset has been used in [5, 11, 9].

- Participants: 72 participants: 30 women and 42 men, from 20 to 35 years old (M = 23.5; SD = 2.1).

- Stimuli: 60 videos belonging to four visual categories: faces, one moving object (one MO), several moving objects (several MO) and landscapes. Videos lasted from 10 s to 24.8 s (M = 16.9; SD = 4.8 s).

- Protocol: Participants had to look freely at the 60 videos: 15 videos with faces, 15 with one MO, 15 with several MO and 15 with landscapes. In each visual category, videos were either displayed with their original soundtrack (OriginalSounds), with the soundtrack from another video belonging to the same visual category (SameCatSounds), or with soundtrack from videos belonging to other visual categories (LandscapesSounds, FacesSounds or MovObjectsSounds). The different auditory conditions were balanced. Note that this experiment differ from the others since it features mutliple audio conditions. However, in [11] we found no difference between non original conditions, and proposed that observers might just filter out the unrelated audio information to focus on the sole visual stream. Thus, non original and visual only conditions could be quite similar.

Apparatus is identical to the first experiment.

2.3. Third Experiment

This dataset has been used in [10].

- Participants: 40 participants: 28 men and 12 women, from 22 to 36 years old.

- Stimuli: 15 videos extracted from the AMI Meeting Corpus (different meetings between four colleagues). Each video lasts between 20 and 80 seconds. The resolution is 1232 x 504 pixels (43.4 x 15.5 degrees). Dialogues are in English.

Protocol and apparatus are identical to the first experiment.

2.4. Fourth Experiment

This dataset was recorded to add general purpose movie-based videos to complement experiments 1 to 3. This is an original dataset added to the others which were already acquired in previous work.

- Participants: 24 participants.

- Stimuli: 23 videos extracted from the Hollywood2 database [12]. The resolution is 320 x 176 pixels. Each video lasts between 4 and 30 seconds (M = 10.4; SD = 6.6 s). The videos are extracted from general movies and on each video there is an event which has characteristic audio signature.

- Apparatus: Participants were seated 60 cm away from a 375 x 300 LCD monitor with a spatial resolution of 1280 x 1024 pixels. The eye-tracker is a binocular Seeing Machines FaceLab 5 system recording at 60Hz.

Protocol is identical to the first experiment.

Overall, the dataset contains 148 videos explored by a total of 176 participants in different audio conditions. The new dataset was split into three visual categories: moving objects, landscapes and faces. This classification allows a more detailed analysis of scores than in previous analysis. Fig. 1 shows three frames from three clips. Each clip belongs to a different category - from left to right: moving objects, landscapes and faces. Red points represent gaze positions of participants in visual-only condition, and green points in audio-visual condition. The presence or absence of sound seems to influence the spatial distributions of eye positions, at least for the moving objects and faces categories. To quantitatively test this hypothesis, we need to define some metrics.

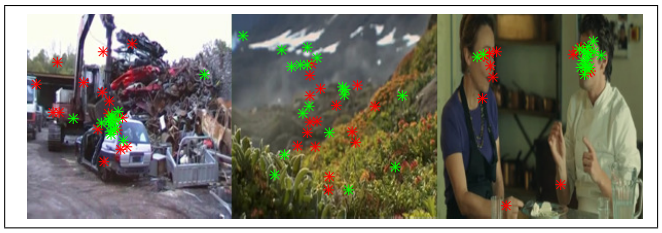

Fig. 1. Three frames of from the moving object, landscape and face categories. The red points represent eye positions of participants in Visual condition, and the green points in AudioVisual condition.

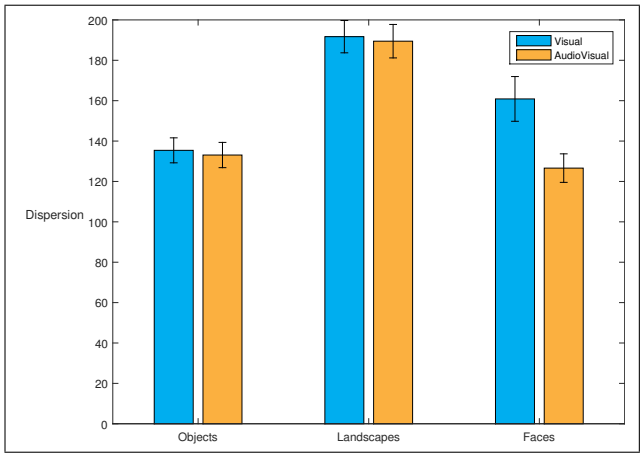

Fig. 2. Mean scores of dispersion by category for visual (in blue) and audio-visual (in yellow) conditions. Lower dispersion means that observers have been focusing on the same locations.

3. METRICS

In order to measure the differences between auditory conditions, five metrics are defined. They also allow spotting temporal locations where audio information particularly matters. They can be grouped in two general categories:

- Inter-metrics, which compare gaze between groups. Eye positions recorded in V condition are compared to eye positions recorded in AV condition.

- Intra-metrics, which are used within a group. Those metrics return two independent values - V condition value and AV condition value.

These metrics can either compare fixation maps (which collect all the observers' eye positions), density maps (fixation maps convolved with a 2D gaussian filter) or both. In this paper, four inter-metrics and one intra-metric will be used.

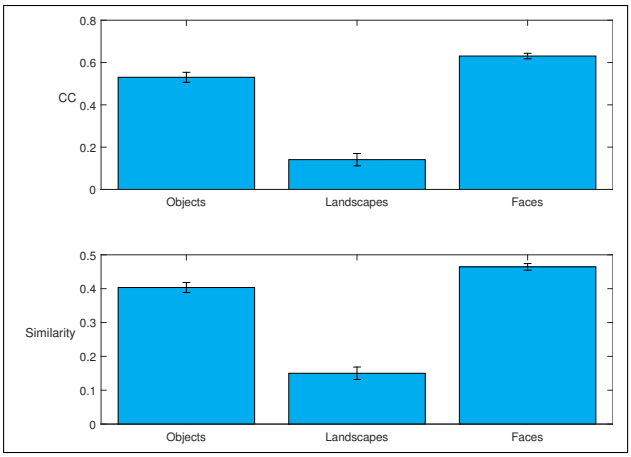

Fig. 3. Mean score of Pearson's Correlation Coefficient (CC) and Similarity for each category. Higher scores show a better agreement between visual and audio-visual conditions while lower scores mean that groups focus on different locations.

3.1. Inter population metrics

These metrics receive as input either density maps or saliency maps (computed from a visual attention model). They can be used for two main purposes. (1) Two density maps computed from eye positions recorded in different audio conditions can be compared to quantify the difference between the latter. (2) A density map can be used as ground-truth to assess the efficiency of a saliency model: the closer the saliency map and the ground-truth, the better the model. Here we use four metrics, compared in [13]:

- the Pearsons Correlation Coefficient (CC), which describes the linear relationship between two variables,

- the Similarity, which represents the summed minimum values between two density maps,

- the Normalized Scanpath Saliency (NSS), which quantifies the saliency map values at the eye fixation locations,

- the Kullback-Leibler Divergence (KL Divergence), which is the loss created when the saliency maps probability distribution is used to approximate the density map.

The KL Divergence is the only dissimilarity metric. The lower the score, the closer the two maps are.

3.2. Intra population metrics

To compare the variance between of eye positions within the same group, we use the dispersion metric [14, 4]. The dispersion is computed for each frame and is defined by:

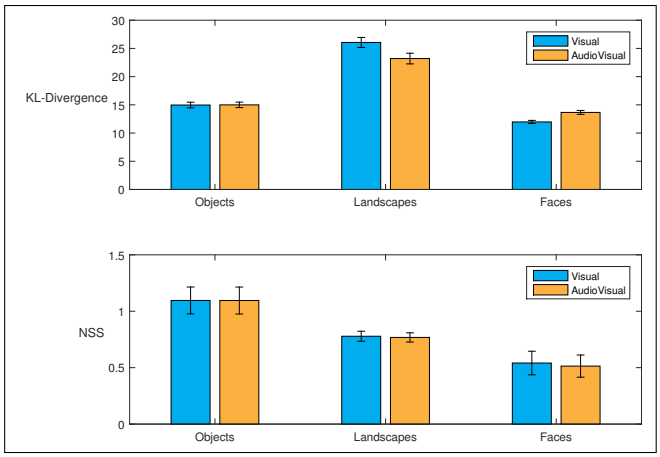

Fig. 4. Mean score of NSS and KL-Divergence of each category. For NSS, higher scores show a better agreement between visual and audio-visual conditions while lower scores mean that groups focus on different locations. For KL-Divergence, lower scores show a better agreement between visual and audio-visual conditions while higher scores mean that groups focus on different locations.

$$D = \frac{1}{N} \sum_{i,j<i} d_{i,j} \quad (1)$$

where $d_{i,j}$ is the Euclidean distance between the eye positions of participants in the same group and N is the number of participants in this group.
Here, we use the dispersion to compare the eye positions within the same population (V or AV condition). The absolute difference between the dispersion computed from the V and AV conditions, is a good marker of the frames where auditory stimuli have a strong impact on participants' attention.

4. RESULTS & ANALYSIS

Fig. 2 shows the mean scores of dispersion by category for visual (in blue) and audiovisual (in yellow) conditions. First, we observe that audiovisual dispersion is always lower than visual dispersion. An interpretation is that audio information drives the viewers toward the same salient objects. Moreover, this trend is much more pronounced in the face category. This observation shows that hearing the original soundtrack makes participants follow the speech turn-taking more closely [11]. These results are in line with Coutrot's PhD thesis [15].

However, the dispersion score only provides a part of the interpretation. Fig. 3 complements the dispersion by providing the Similarity score and Pearson's correlation coefficient. Indeed, those are distribution-based metrics: they compare the spatial eye position distributions rather than the distance between each viewer. The mean score for the Face category is higher than for other categories. It indicates that in visual and audiovisual conditions, observers looked at the same areas in the videos (conversation partners' faces). Moreover, we observe that the similarity and correlation scores on the moving object category are just slightly lower. This suggests that motion attracts the attention of observers whatever the audio condition. Finally, for the landscape category, these metrics are low. This suggests that both groups focus on varied locations: nothing special attracts viewers' attention, who then let their gaze wander over the video. This interpretation is backed-up by the high dispersion recorded for this category.

Fig. 4 confirms the interpretation made in Fig. 3. Those are hybrid metrics: they compare the spatial distribution of the eye position to fixations. We can see that the lowest dispersion is made on Face category and the highest is reached in the Landscape one. However, NSS shows surprisingly different results. This metric represents how the the dispersion of the eye position of one condition encompass the fixation of the others. Object category, by definition, contains objects that, no matter the condition, draws attention. This can explain why the score is more important than the other categories. Moreover, Face category reaches the lowest score on this metric. This reflects the observation made on Fig. 3 and in Coutrot's PhD thesis.

5. DISCUSSION AND CONCLUSION

We propose a new eye-tracking dataset in audio and non audio conditions which gathers several existing datasets and adds an new complementary one. We also propose a Toolbox which has specific metrics to analyze. They are both available only at our website [16]. The presented results illustrate the wide possibilities offered by our dataset and associated ToolBox to compare eye tracking data from two groups of people. This Dataset and ToolBox are meant to evolve, to be complemented by new videos or metrics. In particular, metrics acknowledging that eye movements are a signal that unfold over time, such as the Levenshtein distance, the recurrence quantification analysis or Hidden Markov Models, are destined to grow stronger in eye-tracking studies [17, 18, 19]. The current version of the ToolBox only uses metrics on visual elements to determine when sound impacts on visual attention. However, it does not take into account the auditory or visual features that could also have a significant influence. Future works will focus on integrating various audio / video features extracted from the videos, to determine not only when, but how the soundtrack impacts on participants' attention.

6. REFERENCES

[1] H McGurk and J MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[2] Matei Mancas, *Computational attention towards attentive computers*, Presses univ. de Louvain, 2007.

[3] Andrew J King, "The superior colliculus," *Current Biology*, vol. 14, no. 9, pp. R335–R338, 2004.

[4] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, Alice Caplier, et al., "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, 2012.

[5] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier, "Video viewing: do auditory salient events capture visual attention?," *Annals of Telecommunications*, vol. 69, no. 1, pp. 89–97, 2014.

[6] Guanghan Song, Denis Pellerin, Lionel Granjon, et al., "Sound effect on visual gaze when looking at videos," *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 2034–2038, 2011.

[7] Antoine Coutrot and Nathalie Guyader, "Toward the introduction of auditory information in dynamic visual attention models," in *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[8] Georgios Evangelopoulos, Athanasia Zlatintsi, G Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Y Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 3553–3556.

[9] Antoine Coutrot and Nathalie Guyader, "An Audiovisual Attention Model for Natural Conversation Scenes," in *IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014.

[10] Antoine Coutrot and Nathalie Guyader, "An Efficient Audiovisual Saliency Model to Predict Eye Positions When Looking at Conversations ," in *European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015.

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

[12] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "A study of parameters affecting visual saliency assessment," *arXiv preprint arXiv:1307.5691*, 2013.

[14] Sophie Marat, Tien Ho-Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué, "Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, 2009.

[15] Antoine Coutrot, *Influence du son lors de l'exploration de scenes naturelles dynamiques*, Ph.D. thesis, Université de Grenoble-Alpes, 2014.

[16] Matei Mancas, Julien Leroy, Nicolas Riche, and Pierre Marighetto, "Attention website," http://tcts.fpms.ac.be/attention/.

[17] Li Yujian and Liu Bo, "A normalized levenshtein distance metric," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1091–1095, 2007.

[18] Nicola C Anderson, Walter F Bischof, Kaitlin E W Laidlaw, Evan F Risko, and Alan Kingstone, "Recurrence quantification analysis of eye movements," *Behavior Research Methods*, vol. 45, pp. 842–856, 2013.

[19] Antoine Coutrot, Janet Hsiao, and Antoni Chan, "Scanpath modeling and classification with hidden markov models," *Behavior Research Methods*, vol. in press, 2017.