SOUND EFFECT ON VISUAL GAZE WHEN LOOKING AT VIDEOS

Guanghan Song, Denis Pellerin, Lionel Granjon

Department Images and Signal, GIPSA-Lab
BP 46, 38402 Grenoble Cedex, France
phone: +33(0)476574356, fax:+33(0)476574790, email: guanghan.song@gipsa-lab.grenoble-inp.fr

ABSTRACT

This paper presents an analysis of sound effect on visual gaze when looking at videos to help to predict eye positions. First, an audio-visual experiment was designed with two groups of participants, with audio-visual (AV) and visual (V) conditions to test the sound effect. We classify the sound in three classes: on-screen speech, non-speech and non-sound. We observe with statistical methods that the sound effect is different depending on the class of sound. Then a comparison of the experimental data and a visual saliency model was carried out, which proves that adding sound to video decreases the accuracy of the prediction of the visual saliency model without a sound pathway. Finally, the result of locating the coordinates of a sound source manually provides a viable aspect of sound pathway for future work.

1. INTRODUCTION

In daily life, people do not look at all the objects in the visual field but concentrate on particular regions. These regions that attract attention and therefore, the eyes are called salient regions. Modeling human visual attention which locates salient areas is challenging work. Currently, one of the most influential models of visual attention is the saliency map model proposed by Itti and Koch [4]. The saliency of a spatial area depends mainly on two factors: one is task-independent (bottom-up) and the other is task-dependent (top-down) [3].

In fact, humans usually receive visual and audio information at the same time. Integrating information extracted from audio and video channels is not a trivial task, as it corresponds to different sensor modalities (aural and visual). An earlier application of audio-visual (AV) saliency model is video summarization. Y. Ma et al. [6] and G. Evangelopoulos et al. [2] proposed two different integrations of saliency curves for aural and visual based on different saliency models. Nevertheless, they mainly focused on low-level fusion (at the extracted saliency curves) [10]. This low-level fusion could not contain the information of the saliency region. C. Quigley et al. proposed AV integration research during overt attention, including spatial information [9]. Recently, cross-modal interaction of auditory and visual modalities has played an important role in human spatial saliency and for video coding [5].

Our aim is to study how sound affects visual gaze when looking at videos. Hence, we first describe in Section 2 an audio-visual experiment of two groups of participants with audio-visual (AV) and visual (V) conditions. In Section 3, we analyse the eye positions of the two groups of participants. In Section 4, a comparison of the experimental data and a visual saliency model is presented. In Section 5, we show the pertinence of creating a 'sound localization pathway'.

This research is supported in part by the Rhône-Alpes region (France) with LIMA project.

2. AUDIO-VISUAL EXPERIMENT

The research from C. Quigley [9] showed that sound influences human attention for images. Our purpose is to analyse how sound affects human gaze when looking at videos. In our experiment, the video database was chosen from films which were relevant interesting for both visual and audio. In the visual domain, the database contains various content, including objects, events, characters, sports and so on. In the audio domain, it contains speech, music, noise and some typical sounds, like rain, a knocking door etc.

As we only considered the bottom-up process, the participants viewed the videos without any task. Moreover, in order to reduce top-down effects as much as possible, we created small concatenated clips as proposed in [1]. We put small parts of videos from different sources together with unrelated semantic contents. In order to prevent the participants from understanding the language in the video, we chose foreign languages for the participants, like Chinese, Indian, Japanese etc.

2.1 Stimuli

Sixty video excerpts lasting 5-8 seconds called clip snippets were selected from heterogeneous sources for a total of 10885 frames. Each clip snippet was converted to the same video format (25 fps, 608×272 pixels/frame). The 60 clip snippets were then recombined into 10 clips, each clip being the concatenation of 6 clip snippets from different sources. Because the spatio-temporal model [7] used in section 4 did not consider color information, we used gray level stimuli. Two sets of stimuli were built from these clips, one with AV condition (frames + audio signal), and one with V condition (frames only).

2.2 Participants

Thirty human participants (10 women and 20 men, aged from 21 to 31 years old) were divided to two groups: fifteen participants viewed clips with V condition, and fifteen participants with AV condition. All participants had normal or corrected to normal vision, and reported normal hearing. They were ignorant as to the purpose of the experiment.

2.3 Apparatus and experiment design

Human eye position was tracked by an Eyetracker Eyelink II (SR Research). During the experiment, the participants were sitting in front of a 19-inch color monitor (60 Hz refresh rate) with their chin supported. The viewing distance

between the participant and the monitor was 57 cm. The usable field of view was $20^\circ \times 10^\circ$. The two speakers carried the stereo sound. A 9-point calibration was carried out every five clips. 10 clips are presented to each participant with random order. Before each clip, we present a drift correction, then a fixation in the center of the screen. Participants were asked to look at the 10 clips without any particular task.

2.4 Human eye position density maps

The eye-tracker records eye positions at 250 Hz. We recorded ten eye positions (for the left eye) per frame and per participant. The median of these positions was taken (with X-axis median and Y-axis median) for each frame and for each participant. A 2-D Gaussian was added to each position. The standard deviation of the Gaussian was chosen to have a diameter at mid-height equal to 0.5° of visual angle, which is close to the size of the maximum resolution of the fovea [7]. Therefore, for each frame k and at position (x,y), we obtained a human eye position density map noted $M_h(x,y,k)$.

3. EYE POSITION ANALYSIS

In order to investigate the effect of sound on visual gaze, we considered the eye positions of a group of participants with V condition compared with a group of participants with AV condition. First, we analysed the eye positions of each group of participants separately over time. Second, a comparison of the different eye positions between the two groups was presented.

3.1 Intra each group

With the purpose of evaluating bottom-up influence, we consider the dispersion of the eye positions between the participants in the same group (with AV or V condition). The dispersion D is defined as:

$$D = \frac{1}{N^2} \sum_{i,j<i} d_{i,j}^2 \qquad (1)$$

where, N is the number of participants in one group, $d_{i,j}$ is the Euclidean distance of eye positions between participants i and j.

In Fig. 1, the dispersion is high at the beginning, because eye positions are on the region located at the previous clip snippet. From frame1-9, the dispersion decreases sharply and the minimum value appears at frame 9, because salient regions in a new snippet attract human gaze. The situation is the same in the two groups before frame 70. Subsequently, it is stable in the group with V condition, and increases slowly in the group with AV condition, which means with sound, the regions where the participants looked seem more different over a long period.

3.2 Inter two groups

We analysed the differences of eye positions between the two groups of participants. Fig. 2 is an example of eye positions. In order to measure the distance between the two groups, we adopted the linear correlation coefficient as a criterion, noted as cc. The cc assesses the linearity degree between the two data sets (AV and V conditions). When the cc value is close to -1 or 1, there is an almost perfect linear relationship between the two variables. The cc is defined as follows:

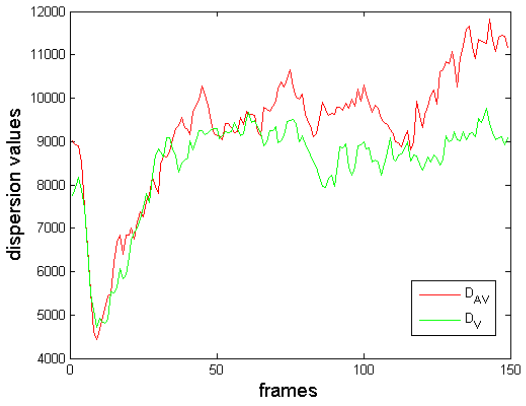

Figure 1: Dispersion D_{AV} (respectively D_V) of eye positions for the group of participants with AV (respectively V) condition as a function of frames. The dispersions are averaged on 60 clip snippets.

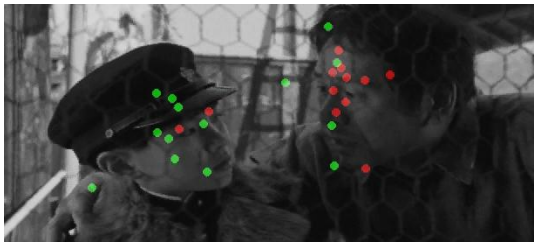

Figure 2: An example of the eye positions of two groups of participants (with AV condition –red points, with V condition –green points)

$$cc(M_{hav}, M_{hv}) = \frac{cov(M_{hav}, M_{hv})}{\sigma_{M_{hav}} \sigma_{M_{hv}}} \qquad (2)$$

where, M_{hav} (respectively M_{hv}) represents the eye position density maps (mentioned in section 2.4), with the AV (respectively V) condition, $cov(M_{hav}, M_{hv})$ is the covariance value between M_{hav} and M_{hv}.

Through observing the video with eye positions, we find that different kinds of sound affect the eye positions differently. Hence, we manually classified the sound into three classes: on-screen speech (the speakers appear on screen), non-speech (any kind of audio signal other than speech) and non-sound (intensity below 40 dB). Because our data is not normally distributed, we used the Kruskal-Wallis test to compare cc in the three classes. The Kruskal-Wallis test is a one-way analysis of variance by rank. It is useful to test equality of population medians among groups.

In Fig. 3, we see that the median of cc increases from on-screen speech to non-speech, finally to non-sound. All the differences are significant, between on-screen speech and non-speech ($\chi^2(1) = 26.91$, $p < 10^{-6}$), between non-speech and non-sound ($\chi^2(1) = 71.67$, $p < 10^{-16}$). The median of on-screen speech is significantly different from the other two classes. It obtains the lowest median value among these three classes with the cc criterion, suggesting the greatest distance between the groups with V and AV conditions.

To ensure the results, we use another criterion named me-

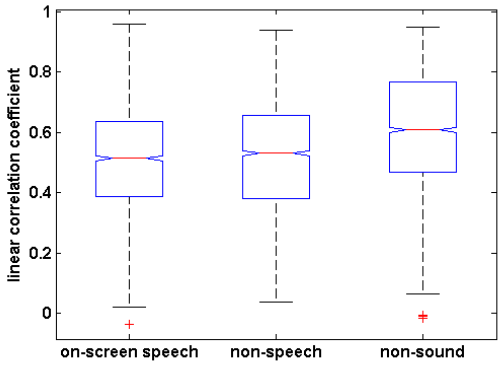

Figure 3: Linear correlation coefficient (cc) between the two groups of participants (with AV and V conditions) in three classes of sound (on-screen speech, non-speech and non-sound).

dian distance (md) to measure the distance between the two groups. It is defined as:

$$md = \frac{1}{N \times N'} \sum_{i \in \mathcal{N}, j \in \mathcal{N}'} d_{i,j} \tag{3}$$

where, $\mathcal{N}$ is the group, with AV condition (number N of participants). $\mathcal{N}'$ is the group, with V condition (number N' of participants). $d_{i,j}$ is the Euclidean distance between participants i and j. Then we used the Kruskal-Wallis test to compare distance in the three classes.

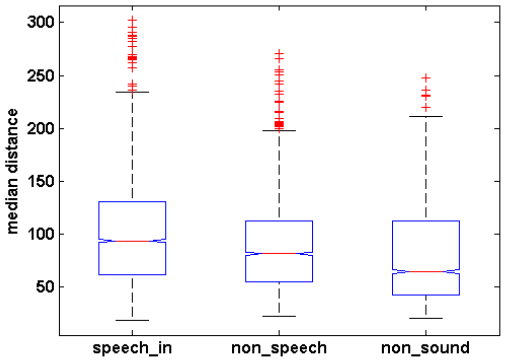

Figure 4: Median distance (md) between the two groups of participants (with AV and V conditions) in three classes of sound (on-screen speech, non-speech and non-sound).

In Fig. 4, the median of md decreases from on-screen speech to non-speech, finally to non-sound. All the differences are significant, between on-screen speech and non-speech ($\chi^2(1) = 77.87$, $p < 10^{-17}$), between non-speech and non-sound ($\chi^2(1) = 36.58$, $p < 10^{-8}$). The median of on-screen speech is significant different from the other two classes. It gets the highest median value among these three classes with median distance criterion, suggesting the highest difference between the groups with V and AV conditions. The trend among these three class is adverse compared to the results from the cc criterion. Nevertheless, the conclusion is the same as cc.

From the above results, we conclude that the eye positions of the two groups (with AV and V conditions) are different when there is sound and the difference is greater when it is on-screen speech rather than non-speech. For the non-sound class, the difference is much lower.

4. EFFECT OF SOUND ON A VISUAL SALIENCY MODEL

In order to test the accuracy of prediction of a visual saliency model for videos with sound, we compare both groups with the spatio-temporal saliency model proposed by S. Marat et al. [7]. This model is inspired by the biology of the first steps of the human visual system. The model extracts two signals from a video stream corresponding to the two main outputs of the retina: parvocellular and magnocellular. Then, both signals are split into elementary feature maps by cortical-like filters. These feature maps are used to form two saliency maps: a static (singularity on one frame) and a dynamic one (motion along two consecutive frames).

For the evaluation, we chose the Normalized Scanpath Saliency (NSS) criterion which was especially designed to compare eye fixations with the salient areas emphasized by a model saliency map [8]. The NSS metric corresponds to a Z-score, which expresses the divergence of experimental results from a model mean as a number of standard deviations of the model. The larger the value of Z is, the less probable it is that the experimental results are due to chance. We computed the NSS metric as follows:

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}} \tag{4}$$

where, $M_h(x,y,k)$ is the human eye position density map standardized to mean 0 and variance 1, and $M_m(x,y,k)$ is the model saliency map. First, we calculated the NSS, successively from dynamic pathway and static pathway, for the two groups. Then we analysed the difference of NSS for each frame between two groups in three classes.

4.1 Dynamic pathway

With the purpose of testing the prediction accuracy of dynamic pathway, we calculated the NSS difference ($NSS_V - NSS_{AV}$) between groups with V and AV conditions in three classes.

In Fig. 5, the median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-19}$. The medians of the non-speech class ($p = 0.12$) and non-sound ($p = 0.06$) are not significantly different from 0. From this, we conclude that the accuracy of prediction from dynamic pathway decreases in a group with AV condition compared to a group with V condition, for the on-screen speech class. (There is no significant difference for the non-speech class and for the non-sound class).

4.2 Static pathway

With the same purpose as dynamic pathway, we calculated the NSS difference ($NSS_V - NSS_{AV}$) from static pathway between groups with V and AV conditions in three classes.

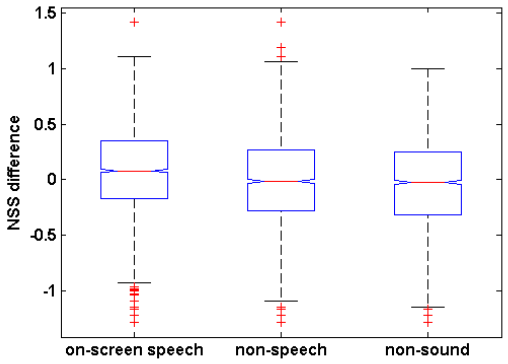

Figure 5: NSS difference ($NSS_V - NSS_{AV}$) between groups with V and AV conditions for dynamic pathway in three classes (on-screen speech, non-speech and non-sound).

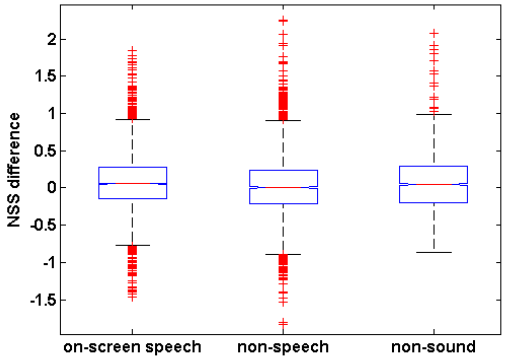

Figure 6: NSS difference ($NSS_V - NSS_{AV}$) between groups with V and AV conditions for static pathway in three classes (on-screen speech, non-speech and non-sound).

In Fig. 6, the median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-32}$. The medians of the non-speech class ($p = 0.10$) and non-sound ($p = 0.05$) are not significantly different from 0. The results are very similar to those obtained from the dynamic pathway and the conclusion is identical. Then, in the case of video with sound, it would be interesting to complete the visual saliency model by a 'sound pathway'.

5. INTEREST OF A 'SOUND LOCALIZATION PATHWAY'

From our observation, the sound source in the video seems to attract human attention. Hence, we located the coordinates of the sound source manually; we call it sound localization pathway. For each frame, we only located one sound source. Then, we applied a Gaussian to this position to obtain a sound map M_{ms}. At last, we compared with NSS the experimental data of the eye positions (groups with AV and V conditions) and the sound maps (M_{ms}).

In Fig. 7, in most of the frames, the NSS of a group with AV condition (mean NSS on all the clips is 2.34) is greater

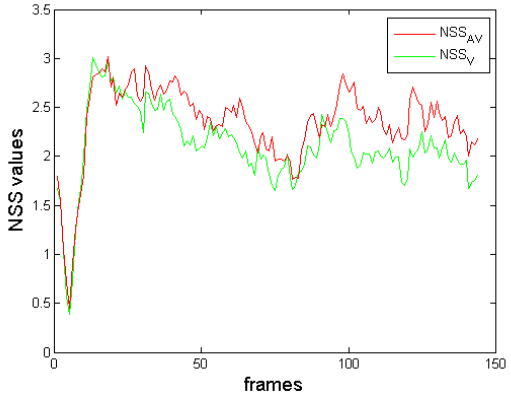

Figure 7: NSS as a function of frame. NSS is averaged on 60 clip snippets of two groups with AV and V conditions for sound map.

than a group with V condition (mean NSS on all the clips is 2.11). Because most of the time the sound source is also the moving or face region on the screen, the group without sound also obtains a high value in this model. Nevertheless, this result shows that locating the sound source is a possible way of increasing the prediction accuracy.

In order to test the prediction accuracy of 'sound localization pathway', we calculated the NSS difference ($NSS_{AV} - NSS_V$) between groups with AV and V conditions in on-screen speech class and non-speech class. Because non-sound class had no sound source in the screen, we did not consider this class.

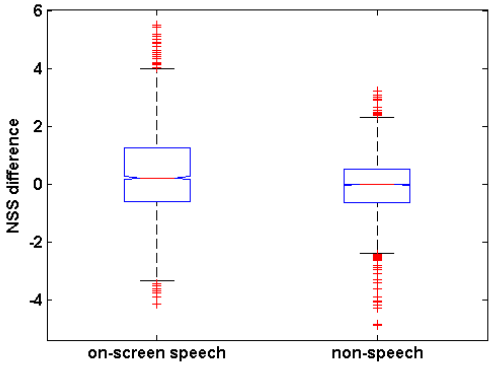

Figure 8: NSS difference ($NSS_{AV} - NSS_V$) between groups with AV and V conditions for 'sound localization pathway' in two classes (on-screen speech and non-speech).

In Fig. 8, the median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-31}$. The median of the non-speech class (p=0.08) is not significantly different from 0. From this, we conclude that the accuracy of prediction from 'sound localization pathway' increases in the group with AV condition compared to the group with V condition, for the on-screen speech class.

6. CONCLUSION AND PERSPECTIVES

This study presents the analysis of the sound effect on human gaze when looking at videos. From our analysis of a group of participants with AV and V conditions, we can conclude that sound affects human gaze differently depending on the sound type, and the effect is bigger for the on-screen speech class. Compared to a visual attention model, the accuracy of prediction decreases in a group with AV condition compared to the group with V condition under the on-screen speech class. If we locate the sound source as a 'sound localization pathway', the prediction accuracy increases a lot. Hence, in future work, it would be interesting to create an audio-visual attention model by adding a sound pathway which locates the sound source automatically. In addition, the sound could be automatically classified to adjust the saliency level.

REFERENCES

[1] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46:4333–4345, Oct. 2006.

[2] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, and et al. Movie summarization based on audiovisual saliency detection. *IEEE Int. Conf. on Image Processing*, 33:2528–2531, Oct. 2008.

[3] J. M. Henderson. Human gaze control during real-world scene perception. *TRENDS in Cognitive Science*, 7:498–504, Nov. 2003.

[4] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, March 2001.

[5] J. S. Lee, F. D. Simone, and T. Ebrahimi. Efficient video coding based on audio-visual focus of attention. *Journal of Visual Communication and Image Representation*, pages 1047–3203, Nov. 2010.

[6] L. Ma, X. Hua, L. Lu, and et al. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, Oct. 2005.

[7] S. Marat, T. H. Phuoc, L. Granjon, and et al. Modelling spatio-temporal saliency to predict gaze direction for short video. *International Journal of Computer Vision*, 82:231–243, May 2009.

[8] R. J. Peters, A. Iyer, L. Itti, and et al. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45:2397–2416, Aug. 2005.

[9] C. Quigley, S. Onat, and S. Harding. Audio-visual integration during overt visual attention. *Journal of Eye Movement Research*, 1:1–17, 2008.

[10] Y. Zheng, G. Zhu, S. Jiang, and et al. Visual-aural attention modeling for talk show video highlight detection. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2213–2216, March 2008.