In Proc. of The 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services, 3-5 July 2013, Paris, France - IEEE copyright

TOWARD THE INTRODUCTION OF AUDITORY INFORMATION IN DYNAMIC VISUAL ATTENTION MODELS

Antoine Coutrot and Nathalie Guyader

Gipsa-lab, Grenoble Image Speech Signal Automatism
CNRS - Grenoble University, France

ABSTRACT

Classical visual attention models only use visual features to predict where observers should look at. However, in daily life, visual information is never perceived without its corresponding audio signal. In a previous study, we found that sound modifies visual exploration by comparing the eye movements recorded when viewing videos with or without their original soundtrack. The aim of the presented research is to further understand how sound influences eye movements by controlling visual and audio contents of the videos, as well as the congruency between them. We describe an experiment with a novel approach in which observers watched videos belonging to four visual categories presenting different visual saliency distributions: landscapes, one moving object, several moving objects and faces. Videos were seen with their original soundtrack or with the soundtrack from another video belonging to the same visual category. Using different metrics to analyze the recorded eye movements, we found that sound has an influence only on videos containing faces and several moving objects. The original soundtrack decreases the variability between the eye positions of observers. Finally, we propose some cues to integrate sound information in classical visual attention models.

1. INTRODUCTION

Humans are able to selectively focus on some parts of an incoming visual scene, discarding less interesting locations. To select the most pertinent parts, the brain uses a filter, called attention. Models of visual attention are mostly based on visual features [1, 2, 3]. A visual input (static image or video frame) is separated into several topographic maps of basic visual features (intensity, color, spatial frequencies, orientations, motion, etc.), which outline the spatial locations that stand out from the background. These maps are merged into a master saliency map that emphasizes the most salient locations of the input. Although in everyday life a visual scene generally comes with a corresponding auditory scene, visual attention models do not take into account auditory information. Yet, it has been shown that synchronous multimodal stimuli are more likely to be further processed, and thus to capture ones attention, than asynchronous or unimodal stimuli [4, 5]. In a previous study, we compared the eye movements of observers viewing videos without sound or with their original monophonic soundtracks. We found that even if fixated areas are similar when looking at video with or without sound, some differences appear. With sound, the eye positions of participants are less dispersed; observers look more away from the screen center, with larger saccades [6]. In the present study, we go further by strictly controlling the visual and the audio contents of the videos. We chose four visual categories: landscapes, one moving object, several moving objects and faces. The last three categories present clustered salient areas (faces and motion being powerful attention attractors). On the contrary, the saliency of landscapes is quite uniform, with no particular region of interest. Each audio source had to be visible. We ran an eye tracking experiment during which observers were asked to freely explore videos with their original soundtrack or with a soundtrack from another video belonging to the same visual category. We hypothesize that a non-original soundtrack might disturb the visual exploration, increasing the variability between observers. We compared the eye positions between the visual categories and the auditory conditions using different metrics commonly used in eye-movement studies. Finally, we discuss some cues to introduce the sound information in dynamic visual saliency models as another feature that, in some cases, plays a role in driving the visual attention.

2. EXPERIMENT

In this section, the eye tracking experiment is described in details. The aim of the experiment was to measure whether different soundtracks modify the gaze of observers when viewing videos with controlled audio-visual content.

2.1. Stimuli

We chose 60 videos belonging to four visual categories: faces, one moving object (one MO), several moving objects (several MO) and landscapes. Face videos mainly present conversations between two people. At any time, every auditory sources had to be visible. When the soundtrack contained

speech, it was always in French. Each video had a resolution of 720 x 576 pixels (30 x 24 degrees of visual angle) and a frame rate of 25 frames per second. They lasted from 10 s to 24.8 s (M = 16.9 $\pm$ 4.8 s). Because it was already shown that shot cut strongly influences eye movements [7, 6], each video was made up of a single shot cut. We chose to focus on the influence of non-spatial sound on eye movements; hence, we used monophonic soundtracks. The video dataset is available online: www.gipsa-lab.fr/~antoine.coutrot/DataSet.html.

2.2. Participants

Twenty students from the University of Grenoble participated in the experiment (*mean age* = 23.5 $\pm$ 2.1). They were not aware of the purpose of the experiment and gave their consent to participate. This study was approved by the local ethics committee. All were French native speakers, had a normal or corrected to normal vision and reported normal hearing.

2.3. Apparatus

Participants were sat 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 head was stabilized with a chin rest, forehead rest and headband. The audio signal was played via headphones (Sennheiser HD 280 Pro, 64 Ω). Eye movements were recorded using an eye tracker (Eyelink 1000, SR Research) with a sampling rate of 1000 Hz and a nominal spatial resolution of 0.01 degree of visual angle. Thus, an eye position was recorded each millisecond in a monocular pupil - corneal reflection tracking mode. Experimental session was preceded by a calibration procedure during which participants had to look at 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.

2.4. Experimental Design

The experiment was designed using a homemade software (SoftEye). This software allows to display the videos and to send to the Eyelink software additional information about the presentation timing. Before each video, a fixation cross was displayed in the center of the screen for 1 second. After that time, and only if the participant looked at the center of the screen (gaze contingent display), the video was played on a mean grey level background. Between two consecutive videos a grey screen was displayed for 1 second. 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. As a whole, the experiment lasted around 20 min. In order to avoid any order effect, videos were randomly displayed. In each visual category, half of the videos were displayed with their original soundtrack (Original condition) and the other half with the soundtrack of another video from the same visual category (Non Original condition). The order of the auditory conditions (Original and Non Original) was randomized. Finally, for each video in each auditory condition, ten observers were recorded.

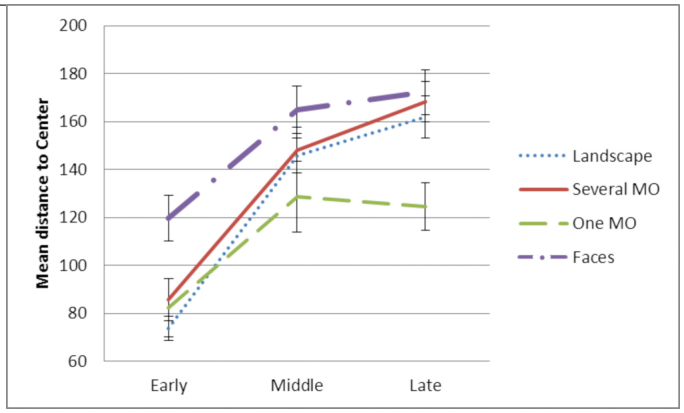

Fig. 1. Mean distance between eye positions and screen center as a function of the visual category (landscape, several moving objects, one moving object and faces) and of the viewing period (early, middle and late). Since this metric is not influenced by sound, results are averaged over the auditory conditions. Distances to Center are given in pixels with error bars corresponding to standard errors.

3. METRICS AND RESULTS

For each frame, we extracted the median eye position of the guiding eye of each subject. Each frame has been seen by 20 persons (10 in the Original and 10 in the Non Original condition). In our analysis, we used three metrics based on these median eye positions and compared them between the visual categories and the auditory conditions. The comparison was done on average over three periods of viewing time: the early period (frames 1 to 25 - first second), the middle period (frame 25 to 75 - second to third second) and the late period (frame 75 to the end of the sequence). Subsequently, when we conclude about differences between conditions we rely on significant differences (ANOVA and Student's tests with p-values less than 0.05). First, we present two metrics that were not influenced by the soundtrack (the distance to center and the number of clusters). Second, we present the results on the dispersion that was influenced by the soundtrack.

3.1. Distance to center

The distance to center is defined as the distance between the barycenter of a set of eye positions and the center of the screen. It expresses the tendency one has to gaze toward the center of the screen while freely exploring a natural scene (central bias) [8]. The soundtrack did not influence this metric.

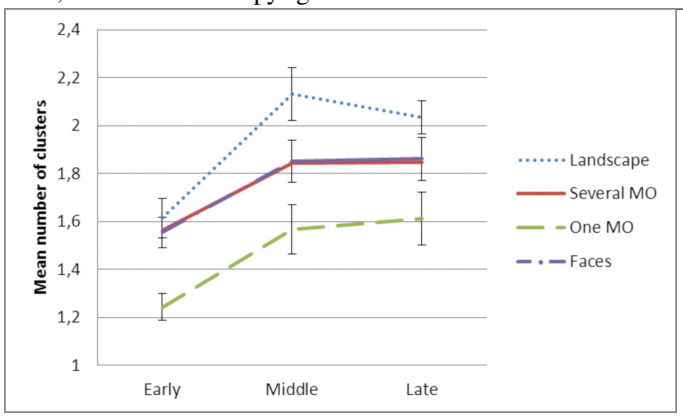

Fig. 2. Mean Number of Clusters as a function of the visual category (landscape, several moving objects, one moving object and faces) and of the viewing period (early, middle and late). Since this metric is not influenced by sound, results are averaged over the auditory conditions. Error bars correspond to standard errors.

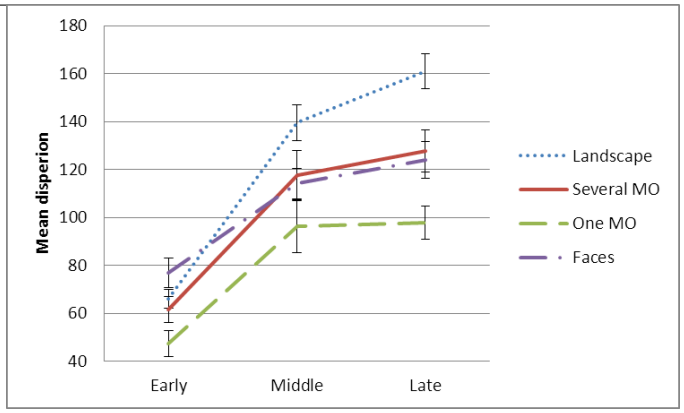

Fig. 3. Mean dispersion as a function of the visual category (landscape, several moving objects, one moving object and faces) and of the viewing period (early, middle and late). Results are averaged over the auditory conditions. Dispersions are given in pixels with error bars corresponding to standard errors.

Figure 1 shows that the mean distance to center is smaller at the beginning of the exploration (early period) compared to the middle and late periods of viewing, independently of the visual category. This result confirms the central bias usually observed during visual exploration. On parallel, we observe that, for faces, the mean distance to center is always larger, from the early to the late period. Indeed, faces are very salient and are rarely at the center of the screen. Moreover, when the video only contains one Moving Object (MO) the distance to center is still short on the late period of viewing. In fact, in our videos, the moving object was mainly at the scene center.

3.2. Number of clusters

We separated the eye movements made on the same video by different participants into clusters via a Mean shift clustering framework. This metric quantifies whether observers are looking at particular locations, called regions of interest, or not. If so, we should obtain a small number of clusters centered on the regions of interest. Mean shift is an efficient clustering algorithm previously used in robust computer vision system and in visual interest quantification [9]. The algorithm considers a given set of eye positions as sampled from the underlying probability density function, and matches each eye position with a local maximum of the density function. As an illustration, Figure 5 shows two faces, on which all the eye positions are clustered. In this particular case, the Mean Shift algorithm returns two clusters. On average and over our videos, the soundtrack did not influence this metric. As shown Figure 2, there is few clusters at the beginning of the exploration (center bias: one cluster at the scene center) and increases across time (middle and late periods). We also found that the landscape category has on average more clusters: since landscapes do not contain particular region of interests, observer's eye positions are spread over the frame. In this case there are several clusters, each one containing only few eye positions. It is the exact contrary of the one MO category where one region of interest (the moving object) focused the attention of all participants.

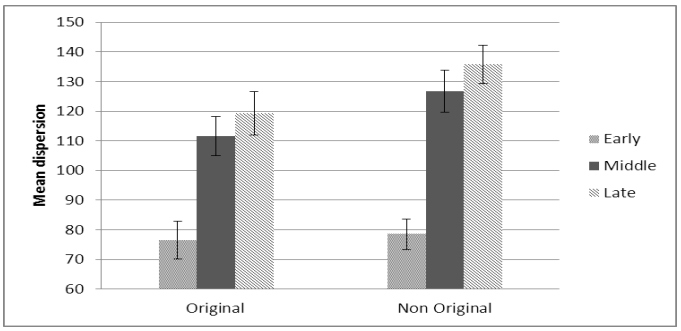

Fig. 4. Mean dispersion as a function of the viewing period (early, middle and late) and of the auditory conditions (Original, Non Original) for the face category. Dispersions are given in pixels with error bars corresponding to standard errors.

3.3. Dispersion

The dispersion estimates the variability of eye positions between observers. It is the mean of the Euclidian distances between the eye positions of different observers for a given frame. In other words, the more participants look at the same location, the smaller the dispersion is. As seen Figure 3, the mean dispersion is small for the four different visual categories during the early period. This is due to the center bias and is in agreement with the small distance to center (Figure

Fig. 5. Frame extracted from a video belonging to the face category. The eye positions of the participants in Original (green) and Non Original (red) auditory conditions have been clustered with the Mean shift algorithm.

1). The dispersion increases with the viewing time. The increase is larger for the landscapes since in this category, no particular visual location attracts observers gaze. On the contrary, the dispersion stays small across the whole sequence for the one MO, in agreement with the small number of clusters (Figure 2). We observe a significant effect of the soundtrack for two visual categories : faces and several MO. Figure 4 shows that, for the face category, the dispersion is higher in the Non Original than in the Original auditory condition. This is in agreement with our previous study that showed that without sound, the variability between observers is greater than with sound [6]. The results for the several MO category are similar (even if the soundtrack effect is smaller).

4. DISCUSSION & CONCLUSION

Previous results suggest that, although visual information leads visual exploration, the gaze of observers viewing natural dynamic scenes is impacted by the original soundtracks, even without spatial auditory information [6]. In the present study, we go further by strictly controlling the visual and the audio contents of the videos. The computed metrics show that sound has an impact on the face and the several moving object categories, with a stronger effect for the face category. A soundtrack from a different video belonging to the same visual category increases the variability between the locations gazed at by different observers after one second, that is for the middle and the late period of viewing. The early period is not impacted since during this phase, the central bias prevails: to quickly get the gist and start exploring the scene, the observers gaze around the center of the screen. It seems natural that the auditory conditions impact on the face and several moving object more than on the one moving object and landscape categories since the auditory information they convey is more relevant. Indeed, speech and the sound of several moving objects are more likely to modulate visual saliency than the predictable noise of a single object or of the wind blowing. These results provide some cues for developing visual attention models taking into account audio information. First of all, visual attention models should take into account the center bias at the beginning of exploration, except for faces, that immediately attract gaze. Then, in the middle and late period of viewing, the audio information should be integrated to modulate the visual saliency of the moving objects and the faces. Audio information modifies neither the visual saliency of landscapes (or static features), nor the visual saliency of a unique moving object on a scene.

5. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] Olivier Le Meur, Patrick Le Callet, and Dominique Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, pp. 2483–2498, 2007.

[3] 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.

[4] Barry E. Stein and M. Alex Meredith, *The Merging of the Senses*, Cambridge, MA: MIT Press, 1993.

[5] Durk Talsma, Daniel Senkowski, Salvador Soto-Faraco, and Marty G. Woldorff, "The multifaceted interplay between attention and multisensory integration," *Trends in Cognitive Sciences*, vol. 14, no. 9, pp. 400–410, 2010.

[6] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[7] Tim J. Smith, Daniel Levin, and James E. Cutting, "A Window on Reality: Perceiving Edited Moving Images," *Current Directions in Psychological Science*, vol. 21, no. 2, pp. 107–113, 2012.

[8] Po-He Tseng, Ran Carmi, Ian G M Cameron, Douglas P. Munoz, and Laurent Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 1–16, 2009.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.