AN EYE-TRACKING DATABASE OF VIDEO ADVERTISING

Lucie Lévêque[1]and Hantao Liu[2]

[1]Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
[2]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

ABSTRACT

Reliably predicting where people look in images and videos remains challenging and requires substantial eye-tracking data to be collected and analysed for various applications. In this paper, we present an eye-tracking study where twenty-eight participants viewed forty still scenes of video advertising. First, we analyse human attentional behaviour based on gaze data. Then, we evaluate to what extent a machine – saliency model – can predict human behaviour. Experimental results show that there is a significant gap between human and machine in visual saliency. The resulting eye-tracking data would benefit the development of saliency models for video advertising or other relevant applications. The eye-tracking data are made publicly available to the research community.

Index Terms— Eye-tracking, visual attention, saliency, video advertising

1. INTRODUCTION

Nowadays, multimedia systems have become an integral part of human activity, including entertainment, education, security and medicine. In many real-world applications, humans rely upon visual media to communicate information or accomplish a task. It is critical to understand how human observers experience visual media, and then use what is learnt to develop useful solutions or tools for improved human experience and automated vision computing systems [1], [2].

Eye tracking – the process of measuring where people look – has been widely used to study how humans interact with visual information and reveal their multimedia experience [3]. For example, eye-tracking is used in radiology to reveal how visual search and recognition tasks are performed, providing information that can improve speed and accuracy of radiological reading [4]. In [5], research is undertaken to investigate how viewers are affected by distortions in images and videos, resulting in more reliable algorithms for visual quality assessment. The methodology of these studies mainly involves the participation of a number of human subjects, recording of eye movements using an eye-tracker, and an agglomerated analysis of the fixation/gaze patterns. For each stimulus presented to a sample of subjects, this gives a topographic representation (i.e., the so-called saliency map) that indicates conspicuousness of scene locations [5]. In a saliency map, the "salient" regions or regions with higher density of fixations designate where the human observers focus their gaze with a higher frequency.

In recent years, there has been a growing interest in the use of eye-tracking technology in the commercial sector or consumer electrics industry, in applications such as web usability, advertising, video gaming and automotive engineering. The eye-tracking data can be statistically analysed and graphically rendered to provide evidence of specific human visual behaviours. This information can be subsequently modelled to assess the effectiveness of a given medium. Ubiquitous Internet access has made online video advertising rise to unprecedented levels [6]. Video advertising is considered to offer informative but "easy to digest" content. Naturally, advertisers must make sure that potential consumers notice and look at the advertised product while experiencing the video content and storytelling. Eye-tracking can be used to find out in what way advertisements should be mixed with the video content and storytelling in order to effectively catch the viewer's eyes. More specifically, eye-tracking data can be collected to quantitatively measure the visibility of a target product relative to the context or storytelling of a video. Knowing this allows researchers to develop advanced computational models that can predict viewers' gaze patterns and, as a result, an advertiser can easily quantify the success of a given advertising campaign without conducting expensive eye-tracking experiments.

In this paper, we perform an eye-tracking experiment using forty still scenes of popular video advertisements. Based on the eye-tracking data, we also evaluate whether the state-of-the-art computational models of visual attention can predict the ground truth.

2. EXPERIMENTAL DETAILS

2.1. Stimuli

Our dataset consists of forty frames that were extracted from forty online video advertisements of diverse content, including categories such as "animation" (i.e., advertisements with computer generated objects), "celebrity" (i.e., advertisements which feature famous people), "indoor" (i.e., advertisements shot in enclosed areas such as a kitchen or a bar) and "outdoor" (i.e., advertisements taken in open places such as a garden or a park). The stimuli were collected on YouTube, from the video advertisement preceding the actual video. There is a wide range of complexity in terms of the spatial position of the advertised product in the video. For example, some videos feature a product closer to the centre of the screen, whereas a product is placed away from the centre in more complex video advertisements. Fig. 1 shows the stimuli used in our experiment. To make a fair comparison, all test images were scaled using MATLAB's *imresize* function using bicubic interpolation to fit our screen resolution of 1080×1920 pixels.

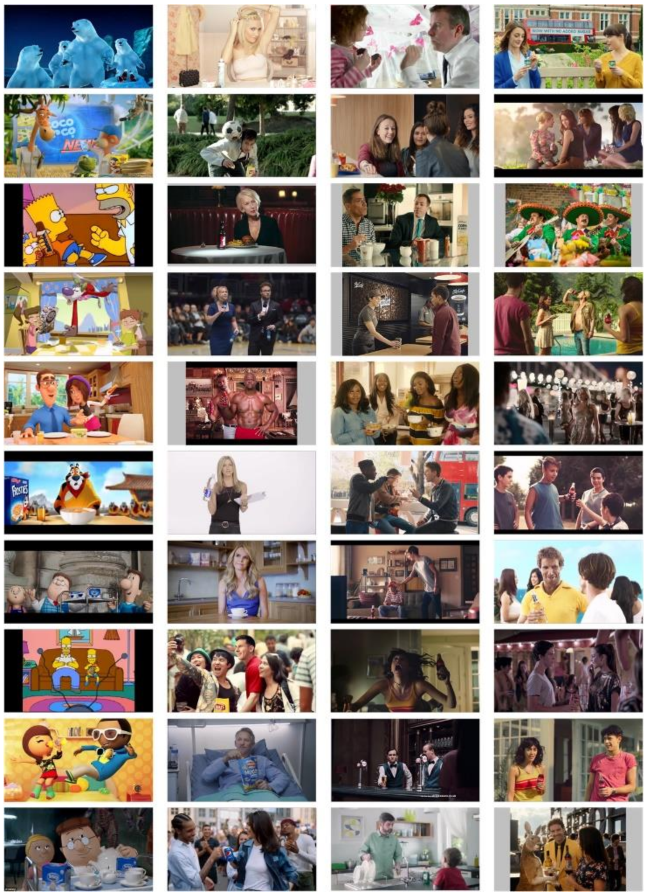

Fig. 1. Illustration of the stimuli used in our experiment. These stimuli were extracted from forty online video advertisements.

2.2. Eye-tracking: experimental procedure

We set up a standard office environment to conduct our eye-tracking experiment [5]. The forty test stimuli were displayed on a 19-inch LCD monitor screen with a native resolution of 1080×1920 pixels. The distance between the participant and the display was maintained approximately between 60 and 65 cm. The eye movements of the observers were recorded using a non-invasive SensoMotoric Instrument (SMI) Red-m advanced eye-tracking device. The system featured a sampling rate of 250 Hz, a spatial resolution of 0.1 degree and a gaze position accuracy of 0.5 degree. Prior to the experiment, the participants were provided with instructions about the procedure of the experiment and, subsequently, a training session to familiarise them with the experiment. The participants were asked to experience the stimuli in a natural way ("view it as you normally would"). Each stimulus was displayed for *one second* and was followed by a mid-grey screen lasting one second as well. The short viewing time was used in order to make the experiment more realistic as users tend to skip the video advertising or not stay on with the video for a long period of time. Stimuli were presented to each subject in a different random order.

A total of twenty-eight participants, fifteen females and thirteen males, from mixed ethnicities, participated in the eye-tracking experiment. Among them, eighteen were university students and ten were professionals. The sample size per stimulus, i.e., twenty-eight participants, is considered adequate as to the evidence published in [7], where research demonstrated that fifteen participants would yield stable or saturated eye-tracking data. The participants were naïve to the purpose of the experiment and had not previously seen the stimuli.

3. RESULTS AND DISCUSSION

3.1. Saliency maps

Fixations were extracted from the raw eye-tracking data using the SMI BeGaze Analysis software package. A fixation was rigorously defined using the dispersal and duration based algorithm and with the minimum fixation duration being 100ms [7]. To render a topographic saliency map for a given stimulus, fixations over all subjects (i.e., twenty-eight in our experiment) are accumulated and each fixation location gives rise to a grey-scale patch that simulates the foveal vision of the human visual system. The activity of the patch is modelled as a Gaussian distribution of which the width approximates the size of the fovea (i.e., two degrees of visual angle) [7]. Fig. 2 shows the saliency map for a sample stimulus.

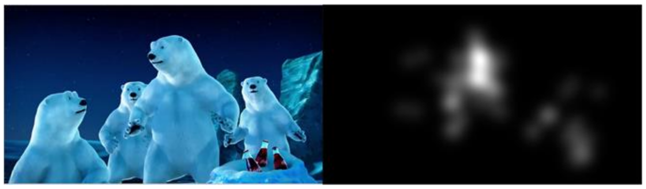

Fig. 2. Illustration of the saliency map for a sample stimulus. The darker the regions, the lower the saliency.

Fig. 3 illustrates the saliency maps obtained for all test stimuli. To better visualise the salient areas, the saliency map, as shown in Fig. 2, is superimposed on top of the original image. The blended saliency maps clearly show, in each case/stimulus, how viewers see the target advertisement while experiencing the video storytelling.

It can be seen from Fig. 3 that in a short (i.e., *one second*) viewing slot, the highly salient regions tend to cluster around visual features that represent storytelling, e.g., the animated characters and their interactions, faces of the celebrities, and humans in active scenes. At the meantime, viewers showed a very good performance in fixating their gaze on target product, independent of its location and size. For example, in some demanding conditions where the target product is far from the centre or hidden in the background, viewers' gaze can be successfully focused on the target.

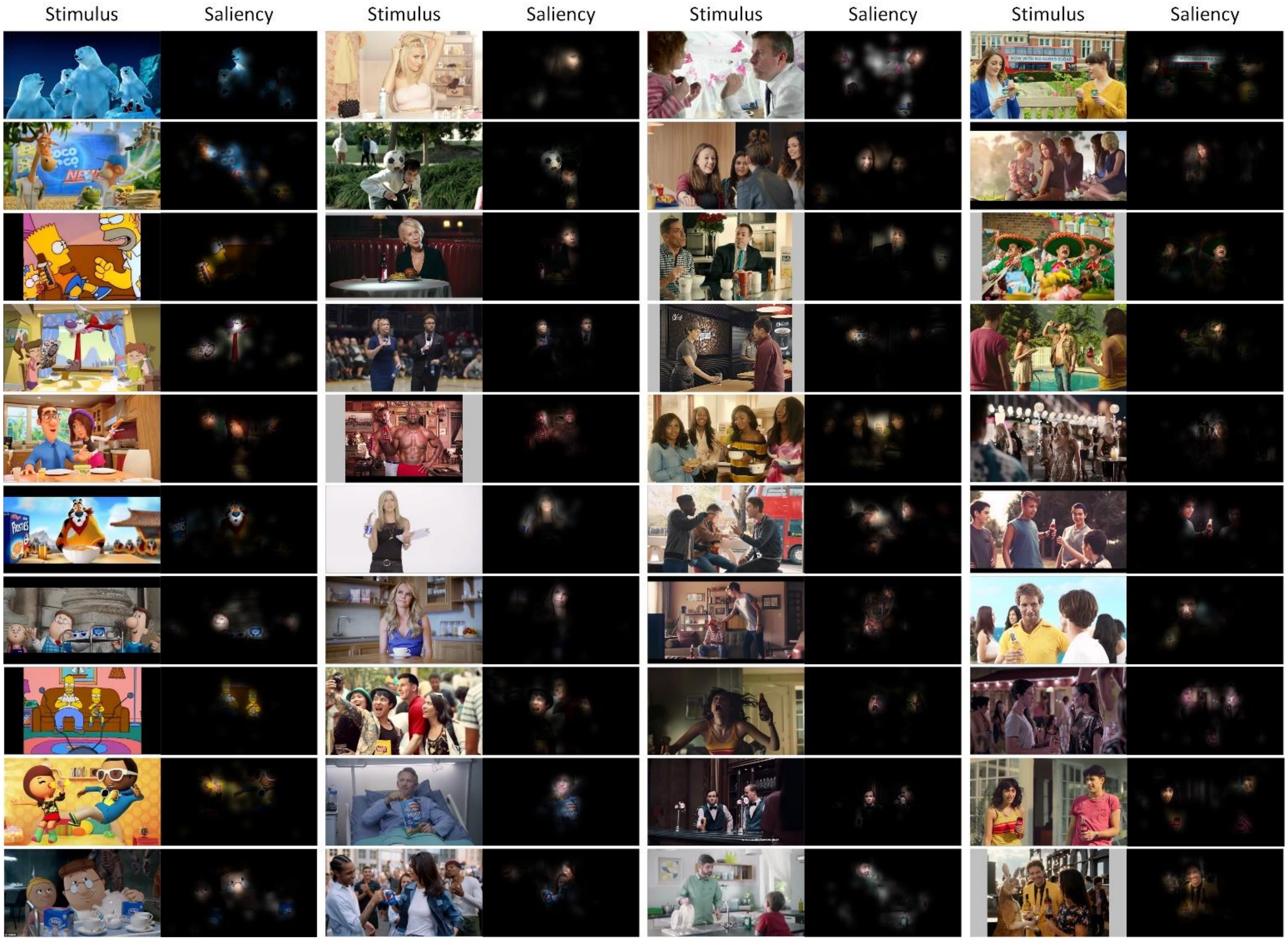

Fig. 3. Illustration of the saliency maps when superimposed to the original stimuli.

3.2. Human vs. machine

Based on eye-tracking data, we have investigated, so far, how human subjects experience video advertising. However, eye-tracking is expensive, cumbersome, and impractical in many circumstances. A more realistic way to use visual attention in multimedia systems is to produce computational saliency. Many saliency models are available in the literature [8]. These models have been developed for different application domains, such as object detection, and, therefore, may reflect different aspects of human attention. To make a saliency model applicable and potentially useful, it is important to validate its prediction accuracy against the ground truth.

We carry out an evaluation with five state-of-the-art saliency models, namely AIM, AWS, GBVS, Itti and RARE2012. AIM [9] is based on the simple principle that attention seeks to the most informative visual content. AWS [10] is grounded on the specific adaptation of low level features. GBVS [11] is a bottom-up visual saliency model composed of the formation and normalisation of activation maps. Itti's model [12] was inspired by the neuronal architecture of the primate visual system. Finally, RARE2012 [13] selects information based on a multi-scale spatial rarity. Fig. 4 shows the computational saliency maps generated by these models for some of the test stimuli in our dataset. It can be seen from the figure that computational saliency models fail in matching with the eye-tracking data. To quantify the similarity between human fixations and a modelled saliency map, three metrics are commonly used, which are as follows: the Pearson linear correlation coefficient (CC), the normalised scanpath saliency (NSS), and the area under the receiver operating characteristic curve (AUC). These metrics are already described in more detail in [8]. In principal, when CC is close to -1 or 1, the similarity is high, whereas when CC is close to 0, the similarity is low. When NSS>0 or AUC>0.5, the similarity measure is significantly better than chance, and the higher the value of the measure the more similar the two variables. Fig. 8 illustrates the similarity measure between human and modelled saliency averaged over all stimuli based on CC, NSS and AUC, respectively. GBVS seems to be the best performing model among five saliency models, however, it shows a poor correlation with human attention. There is still room for improvement in the development of a sophisticated model for the current application.

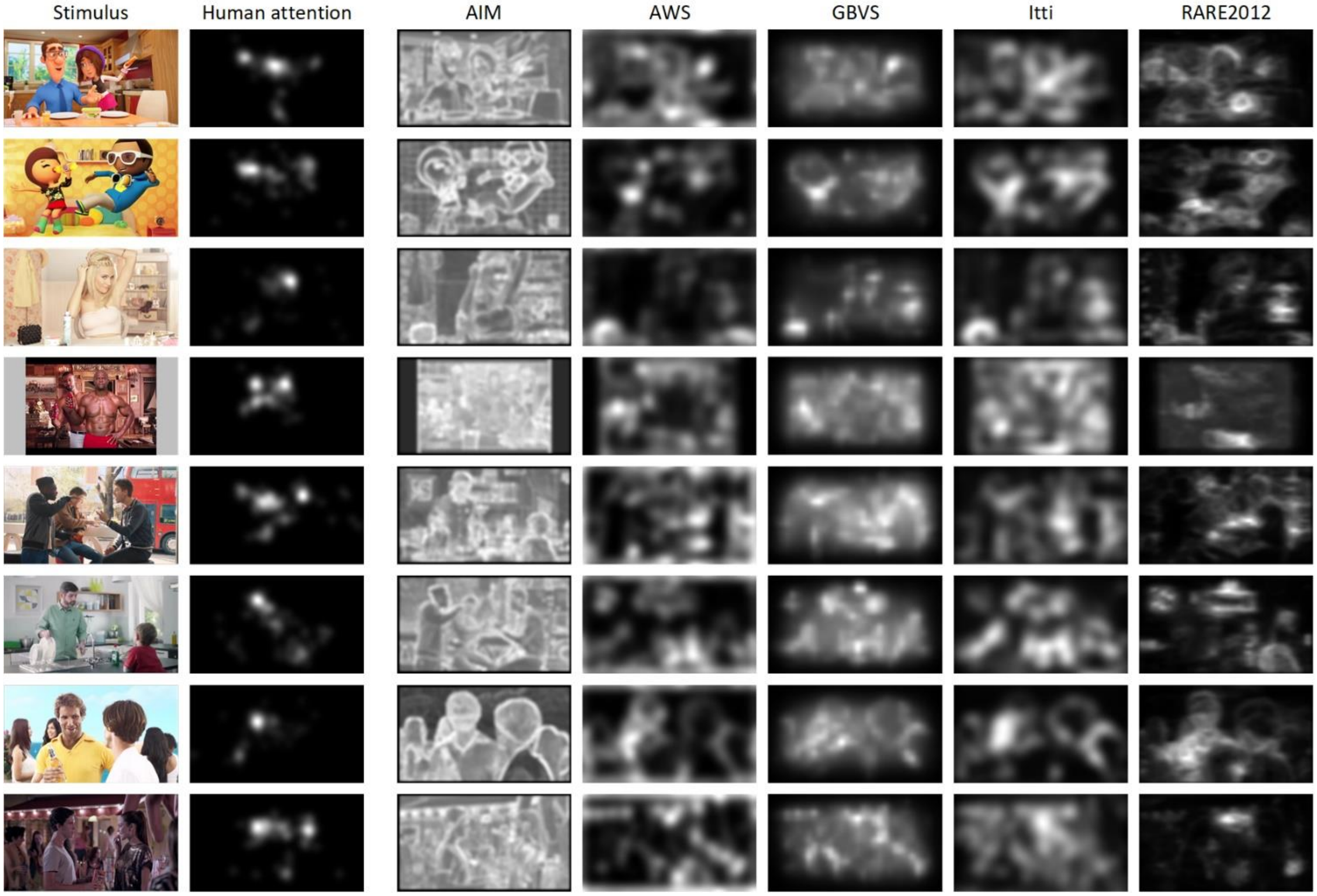

Fig. 4. Illustration of the computational saliency maps generated by five models for some of the test stimuli in our dataset. The second column shows the saliency maps generated from the eye-tracking data. The third to seventh columns represent the saliency maps generated from AIM, AWS, GBVS, Itti and RARE2012 models, respectively.

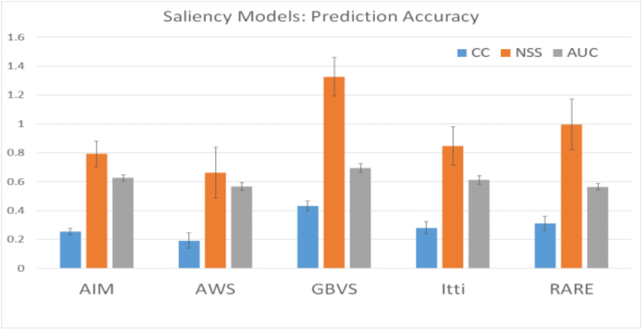

Fig. 5. Illustration of the similarity between human and modelled saliency averaged over the forty stimuli using the CC, NSS, and AUC metrics. The error bars indicate a 95% confidence interval.

4. CONCLUSIONS

In summary, we have used eye-tracking technology to reveal viewers' gaze behaviour in video advertising. In addition, we have assessed whether computational saliency models can be used to replace expensive eye-tracking for this particular application. The results showed a need for improvement in the accuracy of saliency models. The eye-tracking database can be used as a new benchmark of computational modelling of saliency.

5. REFERENCES

[1] H. Liu and I. Heynderickx, "A Simplified Human Vision Model Applied to a Blocking Artifact Metric", CAIP 2007 The 12th International Conference on Computer Analysis of Images and Patterns, Lecture Notes in Computer Science (LNCS), Springer, August 2007.

[2] H. Liu, N. Klomp and I. Heynderickx, "Perceptually relevant ringing region detection method," 2008 16th European Signal Processing Conference, Lausanne, 2008, pp. 1-5.

[3] M. Carrasco, “Visual attention: the past 25 years.”, Vision Research, vol. 51, pp. 1484-1525, 2011.

[4] H. Kundel, C. Nodine, E. Conant, and S. Weinstein, “Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study”, Radiology, vol. 242, pp. 396-402, 2007.

[5] H. Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 7, pp. 971-982, July 2011.

[6] S. Day, “The 9 key advantages to video advertising”, Whiteboard Animation, [Online], http://www.whiteboardanimation.com/blog/the-9-key-advantages-to-video-advertising, retrieved January 2018.

[7] W. Zhang and H. Liu, "Toward a Reliable Collection of Eye-Tracking Data for Image Quality Research: Challenges, Solutions, and Applications," in IEEE Transactions on Image Processing, vol. 26, no. 5, pp. 2424-2437, May 2017.

[8] T. Judd, F. Durand, and A. Torralba, “A Benchmark of Computational Models of Saliency to Predict Human Fixations”, MIT Computer Science and Artificial Intelligence Laboratory Technical Report, 2012.

[9] N. D. B. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” J. Vis., vol. 9, no. 3, p. 5, 2009.

[10] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, “Saliency from hierarchical adaptation through decorrelation and variance normalization,” Image Vis. Comput., vol. 30, no. 1, pp. 51–64, Jan. 2012.

[11] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in Proc. 20th Conf. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2006, pp. 545–552.

[12] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[13] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, M. Gosselin, and B. Dutoit, “RARE2012 : A multi-scale rarity-based saliency deviation with is comparative statistical analysis,” Signal Processing: Image Communication, vol. 28, pp642-658, 2013