TOWARDS A BEHAVIORALLY-VALIDATED COMPUTATIONAL AUDIOVISUAL SALIENCY MODEL

Antigoni Tsiami[1], *Athanasios Katsamanis*[1], *Petros Maragos*[1], *Argiro Vatakis*[2]

[1] School of Electrical and Computer Engineering, National Technical Univ. Athens, Greece
[2] Cognitive Systems Research Institute, Greece

{antsiami, nkatsam, maragos}@cs.ntua.gr, avataki@csri.gr

ABSTRACT

Computational saliency models aim at predicting, in a bottom-up fashion, where human attention is drawn in the presented (visual, auditory or audiovisual) scene and have been proven useful in applications like robotic navigation, image compression and movie summarization. Despite the fact that well-established auditory and visual saliency models have been validated in behavioral experiments, e.g., by means of eye-tracking, there is no established computational audiovisual saliency model validated in the same way. In this work, building on biologically-inspired models of visual and auditory saliency, we present a joint audiovisual saliency model and introduce the validation approach we follow to show that it is compatible with recent findings of psychology and neuroscience regarding multimodal integration and attention. In this direction, we initially focus on the "pip and pop" effect which has been observed in behavioral experiments and indicates that visual search in sequences of cluttered images can be significantly aided by properly timed non-spatial auditory signals presented alongside the target visual stimuli.

Index Terms— audiovisual saliency model, multisensory integration, biologically-inspired, behaviorally-validated

1. INTRODUCTION

Multisensory interaction and integration in the human brain manifest themselves in multiple ways and in multiple contexts [1–4]. Not only our daily experience but also systematic behavioral and neuroimaging experiments reported in the literature provide a considerable amount of evidence that human behavior is effectively influenced by multimodal combinations of perceived sensory information. Multisensory interactions between incongruent sensory streams may lead to illusionary percepts such as the McGurk effect [5], while multi-sensorial percepts which are in agreement often seem to enhance performance in tasks like visual search.

We are particularly interested in the cases when such multisensory effects are linked with the saliency of observed events [6], namely how events draw human attention in a bottom-up fashion. For example, in a visual search task where humans have to identify a target in a heavily cluttered sequence of images in which both the distractors and the target are dynamically changing, it has been observed that reaction times can be lowered significantly when target changes are synchronized with non-localized audio pips because they make the target essentially "pop out" (i.e., become more salient). This is the so-called "pip and pop" effect systematically observed and analyzed in [7]. Similar effects of audiovisual interaction and integration have been the focus of cognitive research for almost two decades now in an effort to understand underlying mechanisms.

In parallel, there has been extensive research focusing on the mechanisms of visual and auditory saliency in isolation. Several findings in these areas have already found their way into computational models that further refine our understanding and allow exploitation of related concepts and interpretations in real applications. The seminal works [8, 9] have set the foundations for developing a visual saliency model that predicts eye fixations during image free-viewing based on image features. The auditory saliency map presented in [10] works analogously in the auditory domain. Building on these well-established, biologically-inspired computational models of visual and auditory saliency, we present an audiovisual saliency model to account for multimodal integration and interaction as these are manifested in behavioral experiments. The purpose of our work is to develop an audiovisual saliency computational model more closely linked to current behavioral findings aspiring to offer insights in human brain function, which in turn may be proven useful in applications as well.

A computational audiovisual saliency model to predict where attention is drawn in an audiovisual scene, i.e., where the eye would be fixated, is for the first time discussed in [11]. It was developed for guiding a humanoid robot. In this model, estimation of visual saliency is based on the Itti et al. approach [9], while for audio, only the spatial properties of the sources are integrated. For a similar application, the model proposed in [12] is based on Bayesian surprise and source localization for auditory saliency map generation and a phase-based approach for visual saliency. In [13] the auditory saliency map is again estimated via source localization and then fused with visual saliency via a product operation. From a different viewpoint, the audiovisual model introduced in [14, 15] for movie summarization and further improved in [16] aims at predicting when, and not where, attention would be drawn in a dynamic scene. All these models are primarily application-oriented and despite having possibly been inspired by cognitive science, no effort has been made to validate their behavior in comparison with behavioral findings. Closer to the nature of our work, Coutrot and Guyader [17, 18] as well as Song [19] have tried to more directly validate their models with humans with their findings indicating that, in movies, eye gaze is attracted by talking faces and music players.

Without making any explicit connection with a particular application and building solidly on experimental human findings, we investigate ways to integrate already well-established models of auditory and visual saliency into a multimodal computational scheme. The proposed scheme will generate an audiovisual saliency map

This research work was supported by the project COGNIMUSE which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources. It was also supported in part by the EU project DIRHA with grant FP7-ICT-2011-7-288121.

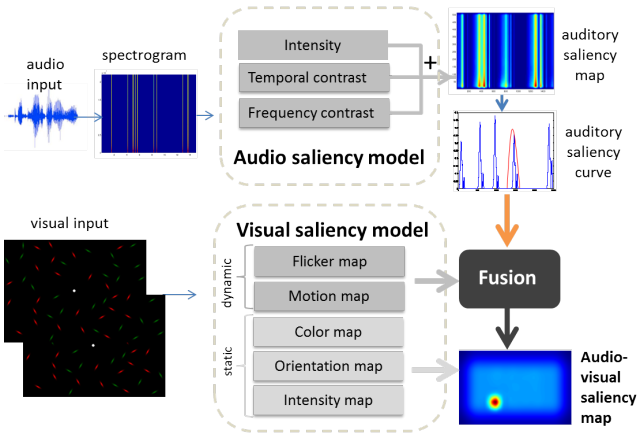

Fig. 1: An overview of the audiovisual saliency model and the individual ones (better viewed in color).

from the input video aiming at accurately reflecting bottom-up multisensory influence to human attention. We initially focus on accounting for the "pip and pop" effect, namely properly fusing the saliency of a sequence of cluttered images with non-localized auditory saliency and we validate the behavior of the model in comparison with corresponding experimental findings regarding humans.

2. AUDIO AND VISUAL SALIENCY MODELS

For visual saliency modeling, the well-known Itti et al. model is employed, the details of which can be found in [9, 20]. We choose this model not only because it is a bottom-up and biologically-inspired model (other options could also be, e.g., [21, 22]), but also because it has been validated with human experiments [6]. More specifically it has been found that model predictions correlate well with actual eye fixation locations, namely the model can reliably predict where human attention is guided. The employed model computes five different types of features. Three of them are static: color, orientation, and intensity and the rest are dynamic: flicker and motion. After the extraction of these low-level features, feature maps are computed at multiple scales. All these feature maps are subsequently normalized, filtered and then summed into the final 2-D visual saliency map. A high level representation of the model is depicted in the lower part of Fig. 1.

Auditory saliency is estimated by means of Kayser et al. model [10]. Analogously to Itti et al.'s model, auditory saliency is estimated on the spectogram image based on three low-level features: intensity, temporal contrast, and frequency contrast. A similar procedure of filtering and normalizing follows feature extraction and leads to a final 2-D saliency map which shows how saliency is distributed over time and frequency (see the upper part of Fig. 1).

3. COMPUTATIONAL AUDIOVISUAL MODELING

Audiovisual integration is a very well-studied manifestation of cross-modal interaction and many behavioral experiments have been carried out to provide insight on how, when, and where auditory and visual information are combined. An example of audiovisual integration in visual search becomes apparent in the presence of synchronized non-spatial audio pips/tones, which lead to lower reaction times behaviorally. This serves as the starting point for the development of our model.

We aspire to appropriately combine the above described individual saliency models in order to form an audiovisual saliency model and investigate its plausibility through comparisons with results from behavioral experiments. In this section the most important issues of audiovisual fusion are discussed, as well as a set of parameters and operators, crucial to the model, that have been found and/or inspired by cognitive research and neuroscience. A high-level overview of the model is presented in Fig. 1.

3.1. From auditory saliency map to auditory saliency curve

As described earlier, the unimodal saliency models generate 2-D saliency maps both for visual and audio streams. Although a 2-D auditory saliency map would be interesting and useful for research on how frequencies contribute to auditory saliency, in this work we are more interested in whether a particular audio event is salient, rather than in how its saliency is related to frequency content. Additionally, as in visual saliency a map denotes which areas of the image are salient and how salient they are, analogously for the audio input, we only need to know if the auditory stimulus is salient and how salient it is.

For these reasons, the 2-D auditory saliency map is further processed to estimate a 1-D auditory saliency curve. The concept of auditory saliency curve is not new [23, 24]. In [24] it was extracted by combining saliencies additively across frequency bins, whereas in [23] the *max* operator was applied for each time instance. Maximization over the entire saliency map was also behaviorally-validated for capturing salient events in [10]. We follow the same approach. With $M_a(t, f)$ we denote the saliency map from Kayser et al. model that is a function of time t and frequency f and with $S_a(t)$ the auditory saliency curve. Thus, the latter is computed as:

$$S_a(t) = \max_f M_a(t, f) \tag{1}$$

The same approach is also followed in [25], where the feature maps are summed and the final temporal saliency score is the maximum for each time instance.

3.2. Audiovisual temporal window of integration

An important finding that originates from cognitive science and has been extensively studied during the last decade is related to the audiovisual temporal window of integration. It has been experimentally found that audiovisual integration is more effective when auditory and visual stimuli are synchronized but can also occur if the two modalities are partially asynchronous. Related behavioral experiments indicate an approximately 200 ms long maximum temporal window of integration [7, 26, 27].

In order to account for this temporal window of integration, auditory saliency is properly filtered. There is also evidence that audition dominates vision in temporal tasks [28–30]. This evidence combined with the fact that audio influences vision even when asynchronous, indicates that we should take into account not only current auditory saliency values, but properly weigh past and future values as well. We employ a Hanning window on the auditory saliency curve, with 200 ms length and center it on the sample corresponding to the current time instance. The form of the Hanning window is suitable for expressing the temporal window of integration because it favors the synchronized stimuli and attenuates the non-synchronized ones. If the stimuli are synchronized, since audio at the current time instance remains unaltered, the integration effect will be maximum, otherwise it will be attenuated to account for the time asynchrony. After windowing, we apply a moving average, thus obtaining a new saliency curve. If we denote by $H(t)$ the Hanning window with N

the window length, and by $A(t)$ the final saliency curve, the latter is computed as:

$$A(t) = \frac{1}{N} \sum_{t_c=t-N/2}^{t+N/2} S_a(t_c) H(t_c) \quad (2)$$

Various other windows of similar properties could be employed, e.g., the Hamming window without any significant differences.

3.3. Audiovisual saliency fusion

The decision of where and how the auditory and visual saliencies will be fused in order to estimate an audiovisual saliency constitutes an important part of our model. In [31] various combination strategies are described to integrate feature maps from different visual-only inputs. These features naturally represent non comparable modalities and they have different dynamic ranges. The same applies also for audio and visual fusion.

Our model is based on the hypothesis that since audio features are dynamic (they constantly evolve), the presence of audio influences mainly the dynamic visual features, flicker, and motion. There would be no effect in fusing audio saliency with the final visual saliency map since it would affect it uniformly.

In parallel, there is evidence that visual flicker and motion are highly influenced by audio given the audio dominance over vision in temporal tasks. Many examples have been presented in the literature [32], such as the bouncing ball illusion [29] which indicate interactions both with flicker [28, 33] and motion [34]. Inspired and motivated by these findings, we fuse auditory saliency with dynamic visual features in order to account for this influence of audio on flicker and motion.

Of particular significance is also the fusion scheme, namely how these different modalities should be fused since they constitute non comparable modalities. In the absence of audio, flicker and motion saliency maps should be left unaltered ,while in the presence of audio its saliency should weigh flicker and motion appropriately. We combine auditory and visual saliencies in a multiplicative manner:

$$F(x, y, t) = F_v(x, y, t)(1 + A(t)) \quad (3)$$

$$M(x, y, t) = M_v(x, y, t)(1 + A(t)) \quad (4)$$

where F and M are the fused flicker and motion maps, F_v and M_v are the visual saliency flicker and motion maps and A is the saliency curve described in the previous section.

This idea has been first presented in a similar but not identical way in [13]. The authors deal with spatial audio only and their auditory saliency map is the location of the audio stimulus. Thus, they combined two 2-D maps with a point-wise multiplication. We extend this idea to fit in our model and data.

4. EVALUATION

In order to quantify our results and validate our model via already published behavioral experiments, we adopted widely used saliency metrics. Since the output of our model is an audiovisual saliency map, where audio saliency has been integrated into the 2-D visual map, we employ metrics commonly used for visual saliency evaluation and particularly those that have been presented extensively in [35, 36].

With Estimated Saliency Map (ESM) we denote the output of our model. Usually, the Ground-truth Saliency Map (GSM) represents the map built from eye movement data. As we currently possess no eye movement data, in our case it represents the ground truth target location in the sense that target is salient, i.e., it "pops out" from the background. Thus, we explicitly create GSM by including the target area, since it is the only salient spot when the target flickers. A similar reasoning was employed also in [6]. The employed metrics are the following (for details see [35, 36]):

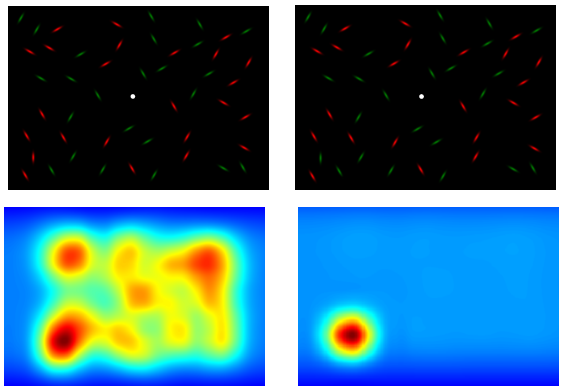

Fig. 2: The two upper figures from [7] depict the "pip & pop" stimuli during a target flicker (the vertical line in the lower left corner). Below are the visual (left) and audiovisual saliency map (right).

1) Normalized Scanpath Saliency (NSS): It is the average of the response values at human eye positions in a model's saliency map (ESM) normalized to zero mean and unit standard deviation. When $NSS \geq 1$, the ESM exhibits significantly higher saliency values at human fixated locations compared to other locations.

2) Linear Correlation Coefficient (CC): It measures the strength of a linear relationship between GSM and ESM.

3) Area Under Curve (AUC): It is the area under Receiver Operating Characteristics curve. Here, target location is considered as the positive set and some points from the image are sampled from the distractors' positions thus obtaining the *shuffled AUC* [35, 37]. The ESM is then treated as a binary classifier to separate the positive samples from the negative ones. The ROC curve is formed by thresholding over the ESM and plotting true positive vs. false positive rate. The area underneath the average of all ROC curves is the *AUC*. When $AUC = 1$ the prediction is perfect.

5. STIMULI AND EXPERIMENTS

For the validation of our model through results from behavioral experiments, we initially only consider simple stimuli from visual search tasks. In such tasks, the participants are instructed to focus at a location on the screen and try to identify a target surrounded by distractors without scanning the image serially. As soon as they identify the target they press a button.

Performance is usually measured by the participant's mean Response Time (RT), i.e. the time from the target appearance to the button press. Mean RT has been linked with saliency in past works, in the sense that it decreases when target saliency increases because it is easier for the participant to identify it [38–40]. We explicitly model this relationship, aiming to explain low mean RT in terms of high saliency and reversely. Thus, our experiments aim to reproduce results and trends from behavioral experiments in relation to saliency, using the metrics described earlier.

The stimuli used in this preliminary validation effort come from visual search tasks [7] and particularly the "pip and pop" effect. They are composed of small straight lines that constantly alter between red and green color. The target is a vertical or horizontal line and when it changes color a non-spatial synchronized audio pip is presented. The rest of the lines have various other orientations and

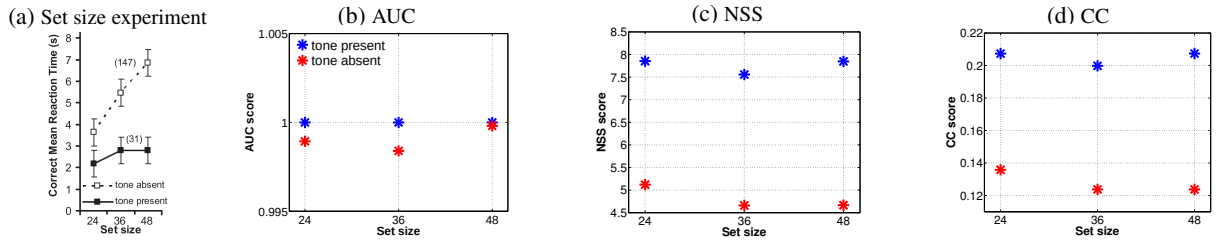

Fig. 3: (a) Original figure from [7] and (b, c, d) *AUC, CC, NSS* for the set size experiment. Blue color depicts the results when tone is present and red color the results when tone is absent.

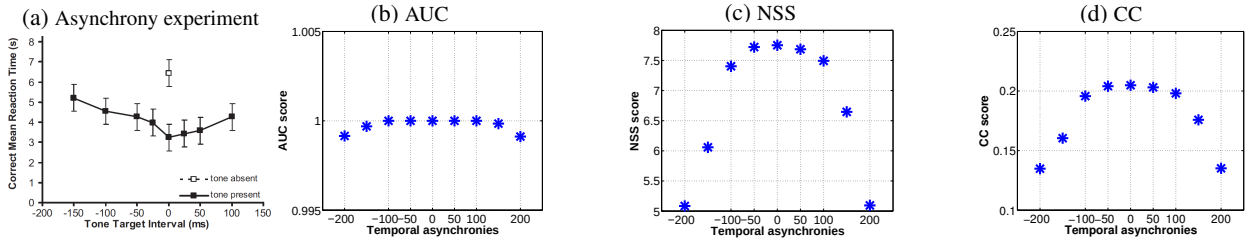

Fig. 4: (a) Original figure from [7] and (b, c, d) *AUC, CC, NSS* for the temporal asynchrony experiment. Minus offsets refer to audio stream preceding the visual one.

their color change is not synchronized with audio pips. In Fig. 2 two stimuli frames are depicted. For the experiments, we have the ground truth stimuli locations as well as the frames when the target changes color, i.e., when it should become salient. Figure 2 also shows the visual-only saliency map and the audiovisual saliency map for the second frame.

5.1. "Pip and pop" set size experiment

In [7] experiments indicate that when there are no audio pips, RTs increase analogously with the number of distractors (target saliency decreases), probably because a serial visual search is required. On the contrary, when a brief synchronized audio pip accompanies the target color flicker, the distractor set size is of no particular importance. The original figure from [7] presenting these results is Fig. 3a.

We investigate whether our model does exhibit this behavior. The input are the "pip and pop" stimuli, the output are the audiovisual saliency maps and the evaluation is carried out with the metrics described in Sec. 4, comparing the audiovisual case with the visual-only one in terms of target saliency. In Fig. 3 we present our results for *AUC, NSS*, and *CC*.

We notice that *NSS* and *CC* reproduce well enough the corresponding Fig. 3a. Although the slopes are not the same, the trend of the curve is similar to the behavioral results. Also, saliency when tone is present is higher than when tone is absent, which is congruent with human data as well. *AUC* is very high for both cases (slightly decreased in the visual-only case) because target flickers alone. It seems that because of the nature of these stimuli this metric cannot capture well the differences between audiovisual and visual saliency. The target in these stimuli is always salient and that is why *AUC* is very high in both cases. *AUC* cannot depict the target saliency increase that is due to the audiovisual integration.

5.2. "Pip and pop" temporal asynchrony experiment

A second behavioral experiment from [7] investigates what happens to audiovisual integration in case the two streams of information are not completely synchronized, namely when salient audio and visual segments that naturally belong to the same event are asynchronous to each other. The findings indicate that audiovisual integration is tolerable in a certain amount of asynchrony. They also depict how asynchrony is related to RTs, showing that the larger the asynchrony is, the more the performance drops and RT increases. These results appear in Fig. 4a. The authors also discuss about the slight asymmetry of the curve in favor of the case when auditory stimulus follows the visual one. We can observe that on the right part of the curve, where the audio pip appears after the target color change, the mean RTs are lower than those of the respective left part, indicating that target becomes more salient when audio follows than when it precedes target flicker. We aim to reproduce this experiment similarly to the previous one. The results are depicted in Fig. 4.

Regarding *AUC* results, we can observe that again target seems to be salient in all cases. Saliency drops slightly for -200 ms and 200 ms asynchrony, but this drop is negligible. However, for *NSS* and *CC* we notice that our results are consistent with the behavioral ones. Saliency increases when synchrony between the two streams increases and vice-versa. The peak of the curve appears when the two streams are completely synchronized. Additionally, target's saliency is higher when audio pip follows target flicker than the opposite, for the same amount of asynchrony, thus exhibiting a similar behavior to the experimental results.

6. CONCLUSION

We have developed a computational audiovisual saliency model based on well-known biologically plausible individual saliency models and aspire to validate its plausibility via human behavioral experiments. Our first validation effort concerns the "pip and pop" effect, where our model exhibits a similar behavior to the experimental results. In the future, we aim to validate our model with other well-known experiments, e.g., [41, 42], and gradually move to more complex stimuli, such as movies.

Acknowledgments

The authors would like to thank Petros Koutras for his helpful remarks and discussions especially on visual saliency.

7. REFERENCES

[1] M. A. Meredith and B. E. Stein, "Interactions among converging sensory inputs in the superior colliculus," *Science*, vol. 221, no. 4608, pp. 389–391, 1983.

[2] M. A. Meredith and B. E. Stein, "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," *Journal of Neurophysiology*, vol. 56, no. 3, pp. 640–662, 1986.

[3] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the unity assumption using audiovisual speech stimuli," *Perception & Psychophysics*, vol. 69, no. 5, pp. 744–756, 2007.

[4] P. Maragos, A. Gros, A. Katsamanis, and G. Papandreou, "Cross-modal integration for performance improving in multimedia: A review," in *Multimodal Processing and Interaction: Audio, Video, Text*. (P. Maragos, A. Potamianos and P. Gros eds), Springer-Verlag, 2008.

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

[6] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[7] E. Van der Burg, C. N. L. Olivers, A. W. Bronkhorst, and J. Theeuwes, "Pip and pop: Nonspatial auditory signals improve spatial visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 34, pp. 1053–1065, 2008.

[8] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

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

[10] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[11] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," in *Proc. ICRA*, 2008.

[12] B. Schauerte, B. Kuhn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *Proc. IROS*, 2011.

[13] S. Ramenahalli, D. R. Mendat, S. Dura-Bernal, E. Culurciello, E. Nieburt, and A. Andreou, "Audio-visual saliency map: Overview, basic models and hardware implementation," in *Proc. CISS*, 2013.

[14] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *Proc. ICASSP*, 2009.

[15] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[16] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, "Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization," in *Proc. ICIP*, 2015.

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

[18] A. Coutrot and N. Guyader, "An audiovisual attention model for natural conversation scenes," in *Proc. ICIP*, 2014.

[19] G. Song, *Effect of sound in videos on gaze: Contribution to audio-visual saliency modeling*, Ph.D. thesis, Universite de Grenoble, 2013.

[20] L. Itti and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. 48th SPIE Int'l Symp. Optical Science and Technology*, 2003, vol. 5200, pp. 64–78.

[21] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," in *Proc. ICASSP*, 2008.

[22] S. Advani, J. Sustersic, K. Irick, and V. Narayanan, "A multi-resolution saliency framework to drive foveation," in *Proc. ICASSP*, 2013.

[23] C. Bordier, F. Puja, and E. Macaluso, "Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging," *Neuroimage*, vol. 67, pp. 213–226, 2013.

[24] O. Kalinli and S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.," in *Proc. INTERSPEECH*, 2007.

[25] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Proc. CISS*, 2012.

[26] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, 2007.

[27] L. Chen and J. Vroomen, "Intersensory binding across space and time: A tutorial review," *Attention, Perception, & Psychophysics*, vol. 75, no. 5, pp. 790–811, 2013.

[28] J. W. Gebhard and G. H. Mowbray, "On discriminating the rate of visual flicker and auditory flutter," *The American Journal of psychology*, vol. 72, no. 4, pp. 521–529, 1959.

[29] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 505–509, 2001.

[30] Y. Wada, N. Kitagawa, and K. Noguchi, "Audio–visual integration in temporal perception," *International journal of psychophysiology*, vol. 50, no. 1, pp. 117–124, 2003.

[31] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[32] L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear.," *Nature*, vol. 408, pp. 788, 2000.

[33] R. B. Welch, L. D. DutionHurt, and D. H. Warren, "Contributions of audition and vision to temporal rate perception," *Perception & Psychophysics*, vol. 39, no. 4, pp. 294–300, 1986.

[34] R. Sekuler, A. B. Sekuler, and R. Lau, "Sound alters visual motion perception.," *Nature*, vol. 385, no. 6614, pp. 308, 1997.

[35] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[36] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[37] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, 2008.

[38] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[39] L. Zhaoping and L. Zhe, "Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data," *PLoS Computational Biology*, 2015 (in press).

[40] H. J. Müller and P. M Rabbitt, "Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 2, pp. 315–330, 1989.

[41] E. Van der Burg, J. Cass, C. N. L. Olivers, J. Theeuwes, and D. Alais, "Efficient visual search from synchronized auditory signals requires transient audiovisual events," *PLoS One*, vol. 5, no. 5, pp. e10664, 2010.

[42] W. Fujisaki, A. Koene, D. Arnold, A. Johnston, and S. Nishida, "Visual search for a target changing in synchrony with an auditory signal," *Proceedings of the Royal Society B: Biological Sciences*, vol. 273, no. 1588, pp. 865–874, 2006.