Audio-Visual Saliency Map: Overview, Basic Models and Hardware Implementation

Sudarshan Ramenahalli[*][†][‡], Daniel R. Mendat[*][‡], Salvador Dura-Bernal[§], Eugenio Culurciello[¶], Ernst Niebur[†][||] and Andreas Andreou[*][‡]

[*]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD
[†]Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, MD
[‡]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD
[§]University of Cyprus
[¶]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN
[||]Department of Neuroscience, Johns Hopkins University, Baltimore, MD

***Abstract*—In this paper we provide an overview of audio-visual saliency map models. In the simplest model, the location of auditory source is modeled as a Gaussian and use different methods of combining the auditory and visual information. We then provide experimental results with applications of simple audio-visual integration models for cognitive scene analysis. We validate the simple audio-visual saliency models with a hardware convolutional network architecture and real data recorded from moving audio-visual objects. The latter system was developed under Torch language by extending the attention.lua (code) and attention.ui (GUI) files that implement Culurciello's visual attention model.**

I. Introduction

Scientists and engineers have traditionally separated the analysis of a multisensory scene into its constituent sensory domains. In this approach, for example, all auditory events are processed separately and independently of visual and somatosensory streams even though the same multisensory event may give rise to those constituent streams. It was previously necessary to compartmentalize the analysis because of the sheer enormity of information as well as the limitations of experimental techniques and computational resources. With recent advances in science and technology, it is now possible to perform integrated analysis of sensory systems including interactions within and across sensory modalities. Such efforts are becoming increasingly common in cellular neurophysiology, imaging and psychophysics studies [1], [2]. A better understanding of interaction, information integration, and complementarity of information across senses may help us build many intelligent algorithms for object detection, object recognition, human activity and gait detection, surveillance, tracking, biometrics *etc*, with better performance, stability and robustness to noise. For example, fusing auditory (voice) and visual (face) features can help improve the performance of speaker identification and face recognition systems [3], [4].

There are several examples of highly successful neuromorphic engineering systems [5], [6] that mimic the function of individual sensory systems. However, the efforts have so far been limited to modeling only individual sensory systems rather than the interaction between them. Our goal in this work is to build computational models of multisensory processing to analyze real world perceptual scenes. We limit our focus to two important sensory systems: the visual and auditory systems. Our work is divided into two parts, one being computational verification and the other being hardware implementation. We investigate the nature of multisensory interaction between the auditory and visual domains. More specifically, we consider the effect of a spatially co-occurring auditory stimulus on the salience of an inconspicuous visual target at the same spatial location among other visual distractors. Temporal concurrency is assumed between visual and auditory events. The motivation for this work is that audio-visual integration is highly effective when cue reliability is highly degraded in respective unisensory modalities. In such a scenario it is beneficial to integrate information from both sensory modalities in order to harness the advantages of each. Neurological studies [7], [2], [8] have shown that audio-visual integration elicits a superadditive response when stimuli in individual modalities are not sufficiently reliable. Results from a hardware implementation of the model are also considered.

II. Related Work

There is a considerable amount of prior research on multisensory processing, specifically audio-visual integration in psychology and neuroscience. For a detailed review of neuroscience and psychophysics research related to audio-visual interaction, please refer to [9], [10]. Here we review research in computational and engineering domains which so far is very limited. We specifically focus on different mechanisms for combining auditory and visual saliency maps. In [11], a one-dimensional computational neural model of saccadic eye movement control by Superior Colliculus (SC) is investigated. The model can generate three different types of saccades: visual, multimodal and planned. It takes into account different coordinate transformations between retinotopic and head-centered coordinate systems, and the model is able to elicit multimodal enhancement and depression that is typically observed in SC neurons [12], [13]. However, the main focus is on Superior Colliculus function rather than studying audio-visual interaction from a salience perspective. In [14], a multimodal bottom-up attentional system consisting of a combined audio-visual salience map and selective attention mechanism is implemented for the humanoid robot iCub. The visual salience map is computed from color, intensity, orientation and motion maps. The auditory salience map consists of the location of the sound source. Both are registered in ego-centric coordinates. The audio-visual salience map is constructed by performing a pointwise *max* operation on visual and auditory maps. In [15], after computing the audio and visual saliency maps, each salient event/proto-object is parameterized by salience value, cluster center (mean location), and covariance matrix (uncertainty in estimating location). The maps are linearly combined based on [16]. Extensions of this approach can be found in [17]. Even though the models in [15], [14] apply audio-visual salience in useful applications, they lack simplicity and biological plausibility. In [18], audiovisual arrays for unthetered spoken interfaces are developed. The arrays localize the direction and distance of an auditory source from the microphone array, visually localize the auditory source, and then direct the microphone beamformer to track the speaker audio-visually. The method is robust to varying illumination and reverberation, and the authors report increased speech recognition accuracy using the AV array compared to non-array based processing.

III. Data and Methods

The effectiveness of audio-visual integration in detecting weakly visible visual target among many distractors is studied by computing an audio-visual (AV) saliency map. The visual stimuli (target and distractors) are deliberately made barely distinguishable from each other. If the auditory stimulus helps identify the target the AV saliency map should reflect the same result. The effectiveness of AV saliency with respect to its unimodal counterparts is studied for different stimulus conditions.

A. Visual Stimuli

The visual stimuli are rectangular imagess with a width of 1800 pixels and height of 150 pixels (Figure 1). A horizontal reference line guides the observer to possible locations of the target and distractors. The task is to identify a weakly visible target symbol among a number of more conspicuous visual distractors in the audio-visual scene. The targets are displayed as the letter ‘Y’ and distractors are displayed as the letter ‘X’ (Figure 1). The number of distractors, D, is randomly chosen to be between 1 and 5 inclusive. There is always *only* one target in every stimulus image. Neither the target nor distractors are allowed to lie within 10 pixels from the image boundaries to avoid unwanted artifacts from the visual salience computation. Distractors are randomly selected without replacement from all possible spatial locations on the abscissa. Among the remaining locations, a target location is randomly chosen. Care is taken to avoid symbols flanking too close to each other. The intensities of both target and distractors are kept identical to avoid intensity-related salience differences. Salience differences in our stimuli are observed because of differences in shape of the symbols only.

Both the distractors and target are distinguishable from the background, but identifying the target from the distractors is a difficult task. If we rely on using the visual domain alone to locate the target, this search requires a considerable amount of attention and thus serial processing to identify if each symbol is the target.

B. Auditory Stimuli

The auditory space is modeled to be spatially coincident with the visual space covering the entire image. We simulate the activation of one of the 8 speakers that are placed equidistant to each other covering the visual space of the entire imgage. So, the auditory space is divided into 8 equal sections. If the visual target (‘Y’) is present in a specific section, a Gaussian window with zero mean and unit variance is centered in that section to represent the approximate auditory signal location. Since auditory localization is generally less precise than visual localization we center the envelope in a particular section of the map irrespective of the exact location of the visual target within that section.

Our model for the auditory signal also serves as an auditory salience map (ASM) because we take spatial location of sound stimulus to be the only relevant feature. Hence, the ASM consists of an activation region if a sound stimulus originates from that location. The sound localization inaccuracy observed in both humans and primates is the motivation to model the stimulus as a Gaussian window (Eq. 1) situated at the location of the sound stimulus:

$$A(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\alpha \frac{(x-\frac{x_o}{2})}{\frac{x_o}{2}}\right)^2} & \text{if } x \in Q_v \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In Eq. 1, Q_v represents the section in which the visual target lies, x_o is the width of the window equal to the length of the active section. The parameter $\alpha = 2.5$, reciprocal of the standard deviation controls the width of the window [19]. The width of the Gaussian roughly corresponds to an uncertainty in auditory stimulus location and the height corresponds to the reliability of the auditory cue.

C. Audio-Visual Saliency

We first compute a proto-object based salience map [20] of the visual scene to investigate the relative visual salience of target and distractors. In the auditory domain, since stimulus location is the only feature considered, the stimulus location map (Figure 3) also serves as the auditory saliency map which is already computed. The visual and auditory saliency maps are combined multiplicatively as:

$$\begin{aligned} S &= f(A) \otimes V \qquad (2) \\ &= (1 + A) \otimes V, \qquad (3) \end{aligned}$$

$$\text{where } V = \frac{1}{3}(n(\bar{O}) + n(\bar{I}) + n(\bar{C})). \tag{4}$$

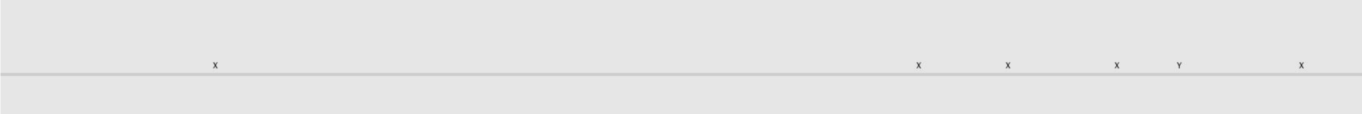

Fig. 1. Visual stimulus with target ('Y') and distractors ('X'). The distractors are visually more conspicuous than target

In Eqs. 2 - 4, S is the audio-visual salience map, A is the auditory salience map, and V is the proto-object based visual salience map. The normalization operator is denoted by $n(.)$, and point-wise multiplication is denoted by the symbol $\otimes$. Color, orientation and intensity conspicuity maps are denoted by $\bar{C}$, $\bar{O}$ and $\bar{I}$, respectively. For more details of visual proto-object based saliency computation, please refer to [20]. By combining the auditory and visual saliency maps as shown in Eq. 4, we retain all salient visual stimuli and also enhance the salience of only those visual stimuli that have a spatially co-occurring salient event in the auditory domain.

D. Hardware Implementation

IV. Results and Discussion

The results of the computational modeling and hardware implementation are presented in Sections IV-A and IV-B, respectively.

A. Computational Modeling Results

In this part, results and discussion regarding the effectiveness of audio-visual integration from a purely computational perspective are provided. The visual proto-object based salience map is computed with default parameters listed in [20]. In the visual domain (Figure 2) we see that distractors are more salient than the target. This salience result implies that an observer is more likely to shift his or her attention to the distractors than to the target. In such a scenario, identifying the target requires an elaborate visual search. On the other hand (see Figure 3), in the auditory domain the section in which the target lies is salient, but the exact location of visual stimulus cannot be identified.

We model the integration of visual and auditory saliencies in a combined audio-visual salience map as described in Eq. 4. The combined audio-visual salience map is shown in Figure 4.

The combined AV salience map illustrates the idea that combining salience maps from multiple modalities can enhance the search for the salient target in an environment among distractors. Despite the fact that the visual salience map makes the target less conspicuous than the distractors, adding the auditory map allows the combined map to roughly identify the location of the stimulus. Without the auditory stimulus, visual distractors exhibiting higher salience than the target are attended to first. However, when an auditory stimulus co-occurs with the target location, the target becomes more salient than the distractors due to multisensory interaction between the auditory and visual modalities. The audio-visual saliency maps for a couple more stimulus conditions are in Figure 5.

Our results confirm the effectiveness of audio-visual salience when cues in unisensory modalities are weak, therefore cannot elicit a strong response based on unisensory cue alone. The effectiveness of multisensory integration is inversely related to effectiveness of unimodal cues [8]. Since we observe increased multisensory salience for the weakly visible target, our model exhibits a form of inverse effectiveness as reported in previous studies [2]. However, the results are preliminary and more testing with different cue reliability conditions is needed to confirm this.

Our model can be advantageous compared to that of [14] because the latter model only highlights salient regions from individual domains. For example, in a scenario where there are three types of events (unimodal auditory, unimodal visual and bimodal audiovisual), the audiovisual event should be more salient than the unimodal events. However, the model from [14] may not account for this. On the other hand, our model assigns higher salience to bimodal events as compared to unimodal ones. Our model also agrees with previous studies [16], [21] where lateralized auditory stimulation was found to topographically increase the salience of the visual field. The model favorably compares with some other experiments where stimulus conditions are slightly different, but visual response enhancement was observed. In [22], a sudden sound, spatially coincident with a *subsequently* occurring visual stimulus was found to improve the detectability of the flash. Our model shows evidence for their main conclusion that involuntary attention to spatially registered sound enhances visual response. In [23] event related potentials were recorded in an audio-visual integration experiment where they found addition of task irrelevant auditory stimulus increased the accuracy and decreased the reaction time in correctly identifying a visual target. It is in agreement with our model.

B. Hardware Implementation Results

V. Conclusion and Future Work

We present a way of combining separate auditory and visual salience maps into an audio-visual salience map where maps from their respective modalities are combined multiplicatively. We retain saliencies of all visual stimuli while enhancing the salience of the target visual stimulus in a model of audio-visual interaction. Without the auditory stimulus, the visual distractors exhibit higher salience compared to the visual target. However, when an auditory stimulus co-occurs with the target visual location the effect is reversed, making the visual target more salient than the distractors. Our results agree with previous neurophysiological studies [2] which establish that audio-visual integration is highly effective when the cue reliability is low in individual modalities taken separately. In

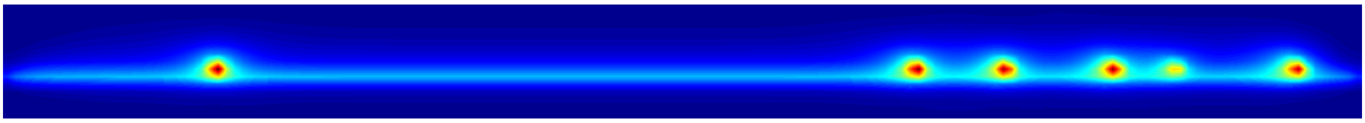

Fig. 2. Proto-object saliency map of the visual stimulus. Notice that target is less salient than distractors

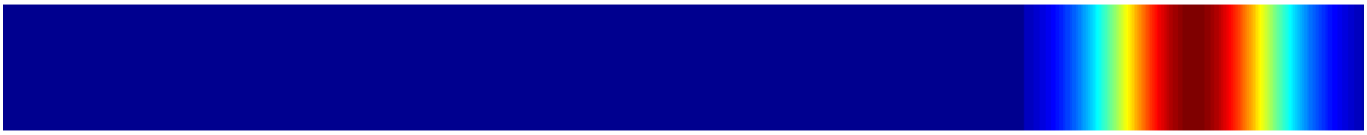

Fig. 3. Auditory stimulus which is also the ASM modeled as a one-dimensional Gaussian. Width of the Gaussian corresponds to uncertainty in location, height to signal reliability

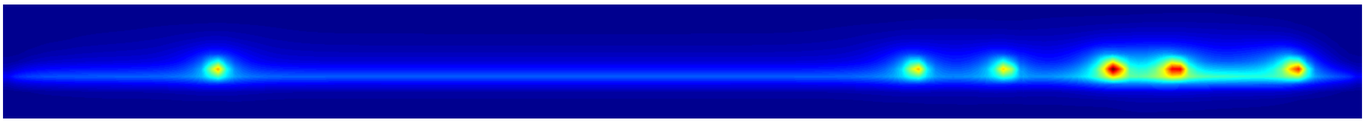

Fig. 4. Combined audio-visual saliency map. Notice the enhancement of saliency of the target. Now, target is more conspicuous compared to the distractors, a reversal effect.

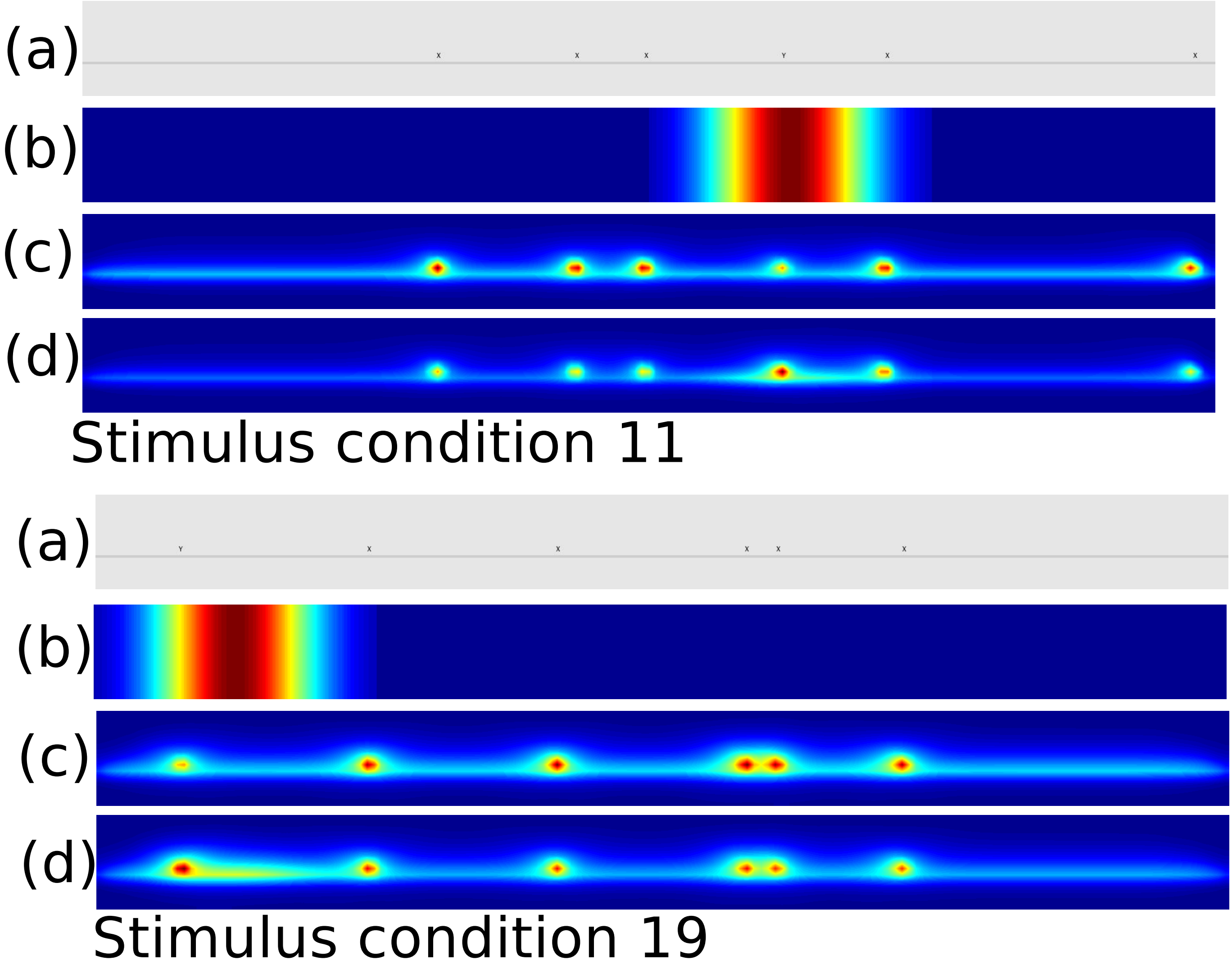

Fig. 5. A few more examples of AV maps: (a) Visual stimuli; (b) Auditory Saliency Map; (c) Visual proto-object based saliency map; (d) Audio-visual saliency map

the future we would like to compare our results with human attention data in multisensory environments.

Acknowledgment

Work supported by Office of Naval Research grant N000141010278 and NIH R01EY016281-02.

References

[1] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 255–266, 2008.

[2] R. A. Stevenson and T. W. James, "Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition," *Neuroimage*, vol. 44, no. 3, pp. 1210–1223, 2009.

[3] H. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio," *Signal processing*, vol. 86, no. 12, pp. 3549–3558, 2006.

[4] S. Tamura, K. Iwano, and S. Furui, "Toward robust multimodal speech recognition," in *Symposium on Large Scale Knowledge Resources (LKR2005)*, 2005, pp. 163–166.

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

[6] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7. IEEE, 1982, pp. 1282–1285.

[7] R. A. Stevenson, M. L. Geoghegan, and T. W. James, "Superadditive bold activation in superior temporal sulcus with threshold non-speech objects," *Experimental brain research*, vol. 179, no. 1, pp. 85–95, 2007.

[8] 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.

[9] D. Alais, F. N. Newell, and P. Mamassian, "Multisensory processing in review: from physiology to behaviour," *Seeing and perceiving*, vol. 23, no. 1, pp. 3–38, 2010.

[10] G. A. Calvert, C. Spence, and B. E. Stein, *The handbook of multisensory processes*. MIT press, 2004.

[11] S. Grossberg, K. Roberts, M. Aguilar, and D. Bullock, "A neural model of multimodal adaptive saccadic eye movement control by superior colliculus," *The Journal of neuroscience*, vol. 17, no. 24, pp. 9706–9725, 1997.

[12] M. A. Meredith and B. E. Stein, "Spatial determinants of multisensory integration in cat superior colliculus neurons," *Journal of Neurophysiology*, vol. 75, no. 5, pp. 1843–1857, 1996.

[13] M. A. Meredith, J. W. Nemitz, and B. E. Stein, "Determinants of multisensory integration in superior colliculus neurons. i. temporal factors," *The Journal of neuroscience*, vol. 7, no. 10, pp. 3215–3229, 1987.

[14] 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 *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 962–967.

[15] B. Schauerte, B. Kuhn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 1173–1179.

[16] S. Onat, K. Libertus, and P. Konig, "Integrating audiovisual information for the control of overt attention," in *Journal of Vision, 7(10):11*, 2007.

[17] B. Kühn, B. Schauerte, R. Stiefelhagen, and K. Kroschel, "A modular audio-visual scene analysis and attention system for humanoid robots," in *Proc. 43rd Int. Symp. Robotics (ISR)*, 2012.

[18] K. Wilson, V. Rangarajan, N. Checka, and T. Darrell, "Audiovisual arrays for untethered spoken interfaces," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ser. ICMI '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 389–. [Online]. Available: http://dx.doi.org/10.1109/ICMI.2002.1167026

[19] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51 – 83, jan. 1978.

[20] A. F. Russell, S. Mihalas, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," submitted.

[21] C. Quigley, S. Onat, S. Harding, M. Cooke, and P. König, "Audio-visual integration during overt visual attention," *Journal of Eye Movement Research*, vol. 1, no. 2, pp. 1–17, 2008.

[22] J. J. McDonald, W. A. Teder-Sälejärvi, and S. A. Hillyard, "Involuntary orienting to sound improves visual perception," *Nature*, vol. 407, no. 6806, pp. 906–908, 2000.

[23] J. Yang, Q. Li, Y. Gao, and J. Wu, "Task-irrelevant auditory stimuli affect audiovisual integration in a visual attention task: Evidence from event-related potentials," in *Complex Medical Engineering (CME), 2011 IEEE/ICME International Conference on*, may 2011, pp. 248 –253.