Visual Attention Driven by Auditory Cues

Selecting Visual Features in Synchronization with Attracting Auditory Events

Jiro Nakajima[1], Akisato Kimura[2], Akihiro Sugimoto[3], and Kunio Kashino[2]

[1] Chiba University
nakajima13@chiba-u.jp
[2] Communication Science Laboratories, NTT Corporation
akisato@ieee.org,kashino.kunio@lab.ntt.co.jp
[3] National Institute of Informatics
sugimoto@nii.ac.jp

Abstract. Human visual attention can be modulated not only by visual stimuli but also by ones from other modalities such as audition. Hence, incorporating auditory information into a human visual attention model would be a key issue for building more sophisticated models. However, the way of integrating multiple pieces of information arising from audio-visual domains still remains a challenging problem. This paper proposes a novel computational model of human visual attention driven by auditory cues. Founded on the Bayesian surprise model that is considered to be promising in the literature, our model uses surprising auditory events to serve as a clue for selecting synchronized visual features and then emphasizes the selected features to form the final surprise map. Our approach to audio-visual integration focuses on using effective visual features alone but not all available features for simulating visual attention with the help of auditory information. Experiments using several video clips show that our proposed model can better simulate eye movements of human subjects than other existing models in spite that our model uses a smaller number of visual features.

Keywords: Visual attention, auditory cues, Bayesian surprise, synchronization, feature selection.

1 Introduction

Human beings have capability of detecting various kinds of objects without any thought or effort. *Visual attention* is considered to play a significant role in achieving this function. In fact, visual attention is one of the built-in mechanisms of the human visual system that quickly selects regions most likely to attract human interest in a visual scene. Such a pre-selection mechanism focusing only on relevant data would be essential in enabling computers to undertake subsequent processing such as generic object recognition or scene understanding.

With this background, many researches have been reported to simulate visual attention in several research fields including psychophysics, neuroscience and computer vision (see extensive survey papers, e.g. [2,3,11] for details). These researches usually

X. He et al. (Eds.): MMM 2015, Part II, LNCS 8936, pp. 74–86, 2015.

take a *bottom-up* approach, meaning that a given video signal is the only resource for simulating visual attention. Nevertheless, they have enabled us to investigate in detail the process of visual search and simulate its performance.

A large amount of effort for developing computational models of human visual attention has ever been devoted to only *visual* processing. Human visual attention, however, can be easily modulated by other modalities. As an intuitive example, when we hear something interest or strange we tend to look at the direction of sounds even if that direction is not so visually salient. As such, sounds are often strongly related to events that draw human visual attention. We will be able to further augment computational models of human visual attention if we incorporate auditory information into them. However, the way of integrating information arising from both audio and visual domains still remains a challenging problem.

This paper proposes a novel model of human visual attention driven by auditory cues. In our model, auditory information plays a supportive role in simulating visual attention, in contrast to standard multi-modal fusion approaches [21,18,19,5,14]. More concretely, we take an approach that detects visual features in synchronization with surprising auditory events. Our strategy is built on two recent psychophysical studies:

1. Audio-visual temporal alignment leads to benefits in visual attention if changes in the component signals are both *synchronized* and *transient* [4].
2. Auditory attention *modulates* visual attention in a *feature-specific* manner [1].

Following these findings, our model first detects *transient* events using the Bayesian surprise model in visual [8] and auditory [20] domains separately, and then looks for visual features in *synchronization* with detected auditory events. Surprise maps are then *modulated* by the selected features, in a similar manner to the guided search [23], one of well-founded psychophysical models that explicitly implements characteristics of target stimuli.

2 Related Work

Building computational models of human visual attention has attracted much attention especially in the last decade. Here we briefly review just a couple of related studies due to the space limitation. Extensive surveys can be found in e.g. [2,3,11] which include the history, detailed taxonomies and related psychophysical findings.

A seminal work as regards bottom-up models of human visual attention is the *saliency map* model proposed by Itti, Koch and Niebur [9]. In this model, the concept of *saliency* as a measure of attractiveness of human visual attention was first introduced into a computational model. Since it is simple, easy to implement and produces reasonable output for various kinds of images, it has had a considerable impact on broader research areas such as image processing, pattern recognition, computer vision, robotics and neuroscience [3].

The saliency map model has been further extended by Itti and Baldi to develop the *Bayesian surprise* model [8] that incorporates temporal dynamics of the human visual system. In this model, saliency is formulated by the Kullback-Leibler divergence between probabilistic density functions (PDFs) of expected and obtained visual features.

Therefore, continuously similar visual features give low saliency values, while unexpected visual features such as sudden changes provide high saliency values. Bayesian inference methods have been introduced also in several other computational models [24,16,6,13]. Such a probabilistic model enables us to handle various types of features with different characteristics into a unified framework. This is why we adopt the Bayesian surprise model as the basis of our new model.

Meanwhile, several mechanisms developed in visual attention models have also been introduced into auditory attention models, for example, the saliency map model [10], SUN (Saliency Using Natural statistics) [22] and Bayesian surprise [20].

However, human visual attention models with the help of auditory information has not been well studied. This might be because solid psychophysical findings about characteristics of audio information in human visual perception have been recently developed. In turn, most existing methods took multi-modal fusion approaches and concentrated on improving application performances, and thus the compatibility with the human visual perception is rather out of focus. Video summarization [12,15,5] is one of the popular applications of audio-visual saliency. Robotics [21,18,19] has also been an attracting application for last several years.

To the best of our knowledge, this is the first work that explicitly incorporates solid psychophysical findings of auditory-based attention modulation into a computational model of human visual attention.

3 Proposed Model

3.1 Framework

Figure 1 depicts the framework of the proposed model. As shown in this figure, our proposed model consists of four main steps.

(1) Bayesian surprise. The first step extracts surprising events in visual and auditory domains individually where image and audio signals are separately applied to the Bayesian surprise model. For a given input video, 360 visual surprise maps with different types of features and a single auditory surprise signal are extracted. The details will be described in Section 3.2.

(2) Synchronization detector. The second step evaluates synchronization of each visual surprise map with the auditory surprise signal. For this purpose, synchronization detectors are attached to every location in each of the 360 visual surprise maps and the auditory surprise signals, resulting in 360 maps. Every map is averaged over pixels to create a sequence describing how synchronized the corresponding visual surprise map is with the auditory surprise. The details will be described in Section 3.3.

(3) Features selection. The third step is devoted to selecting visual features that well synchronize with the auditory surprise. Counting the number of samples with a sufficient level of synchronization for every sequence, we obtain a histogram representing the degree of synchronization for every visual surprise map with the auditory surprise. Remembering that every visual surprise map corresponds to a specific type of features, feature selection based on audio-visual synchronization can be implemented by binarizing the histogram. The details will be described in Section 3.4.

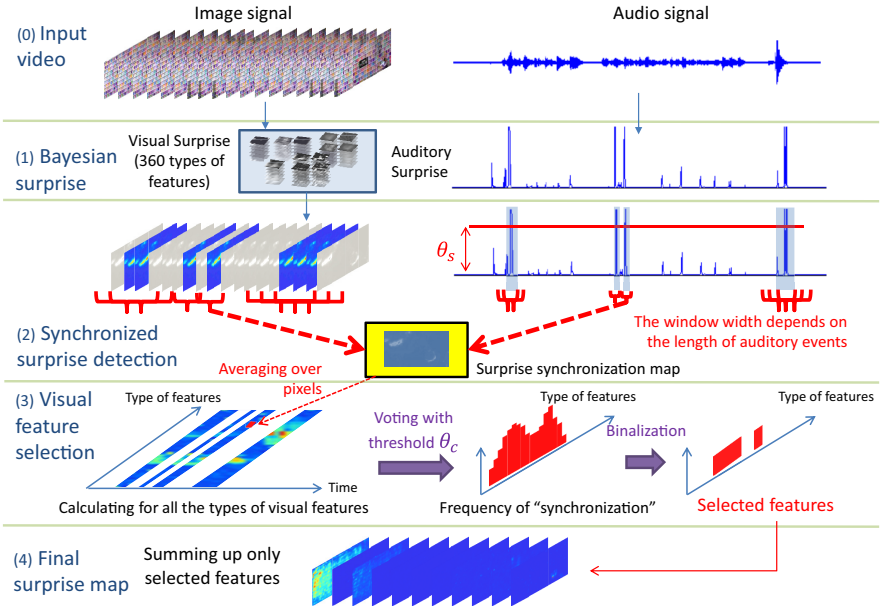


Fig. 1. Framework of the proposed model

(4) Final surprise map. The last step is for forming the final surprise map composed of visual surprise maps with the selected visual features. The details will be described in Section 3.4.

Our proposed model detects transient auditory events, and then selects visual features in synchronization with detected auditory events to modulate the final saliency maps. Note that the proposed model is built on a two-pass algorithm, where the first 3 steps are devoted to selecting visual features that describe major audio-visual events in the input video to produce the final map in the last step.

3.2 Bayesian Surprise Model

Here, we briefly review the Bayesian surprise model proposed by Itti and Baldi [8]. We introduce it to obtain visual surprise maps.

Center-surround feature maps are first generated. They are extracted in parallel over 12 feature channels (intensity, 2 color opponents, 4 orientations, temporal onset and 4 directed motion energies) and 6 spatial scales, yielding $12 \times 6 = 72$ feature maps in total.

Local surprise detectors are then attached to every location in each of the 72 feature maps. Suppose that every pixel value $f(t, \boldsymbol{x})$ received from feature map f at location $\boldsymbol{x}$ and time t obeys a Poisson distribution and it holds a conjugate Gamma prior $\gamma(\cdot\,; \alpha, \beta)$ with parameters (α, β). Once $f(t, \boldsymbol{x})$ is observed, the posterior $\gamma(\cdot\,; \alpha', \beta')$ can be obtained using the Bayes rule. Namely, at each time step t, the posterior $\gamma(f(t-1, \boldsymbol{x}); \alpha_V(t-1, \boldsymbol{x}), \beta_V(t-1, \boldsymbol{x}))$ at the previous step can be used as a prior to obtain

the current posterior $\gamma(f(t,\boldsymbol{x});\alpha_V(t,\boldsymbol{x}),\beta_V(t,\boldsymbol{x}))$. In addition, 5 cascade detectors are implemented at every pixel in every feature map so that the model can detect surprises at several temporal scales. In summary, the update rule of parameters (α,β) at feature map f, time t and cascade level d is described as follows:

$$\alpha_V(t;d) = \xi\alpha_V(t-1;d) + \alpha_V(t;d-1)/\beta_V(t;d-1),$$
$$\alpha_V(t;0) = f(t), \quad \beta_V(t;d) = \xi\beta_V(t-1;d) + 1,$$

where $0 < \xi < 1$ is a forgetting factor and indices f and $\boldsymbol{x}$ are omitted for simplicity.

Local temporal surprise $S_{V,T}(t;f,d)$ at feature map f, time t and cascade level d is determined as the Kullback-Leibler (KL) divergence between the prior and posterior, while spatial surprise $S_{V,S}(t;f,d)$ is as that between the neighborhood prior (modeled as a weighted sum of distributions over neighborhoods at the previous cascade level) and the posterior. The total visual surprise $S_V(t;f,d)$ is determined according to the original paper [8] as

$$S_V(t;f,d) = (S_{V,T}(t;f,d) + S_{V,S}(t;f,d)/20)^{1/3}.$$

As we see, we have in total 72(feature maps) $\times$5(cascade levels) $= 360$ visual surprise maps.

Auditory surprise is derived in a similar manner [20], where a spectrogram $F(t,\omega)$ extracted via short-time Fourier transform (STFT) is used as an observation. Following the same update rule as the visual surprise, we can obtain parameters (α,β) of the posterior at time t and frequency ω as

$$\alpha_A(t;\omega) = \xi\alpha_A(t-1;\omega) + F(t,\omega), \quad \beta_A(t;\omega) = \xi\beta_A(t-1;\omega) + 1.$$

Auditory surprise $S_A(t;\omega)$ at time t and frequency ω is determined as the KL divergence between the prior and posterior. The final auditory surprise $S_A(t)$ is obtained as the mean over all the frequencies. As a result, we have a single auditory surprise signal.

3.3 Synchronization Detector

Following the recent psychophysical insight that audio-visual temporal alignment affects visual attention if changes of component signals are synchronized and transient [4], our model detects synchronized audio-visual events in videos from the output of Bayesian surprise models. Since both audio and visual signals have been converted into the "surprise" domain under the same logic, we can adopt a simple approach based on cross correlation.

A synchronization detector comprises the following 3 steps: Detecting surprising auditory events, pixel-wise cross correlations, and averaging over frames.

(1) Segments of surprising auditory events are first extracted from the auditory surprise signal $S_A(t)$. We exploit a simple approach that extracts segments with a surprise value $S_A(t)$ greater than a predefined threshold θ_s, resulting a set of segments $T_{S,i}$ $(i,1,2,\ldots)$.

(2) For every segment $T_{S,i}$ normalized cross correlation (NCC) is calculated between the auditory surprise signal $S_A(t)$ and visual surprises at every location $\boldsymbol{x}$ in each of the

Table 1. Details of video clips

	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
clip name	advert_bbc4_bees	advert_bbc4_library	sports_kendo	basketball_of_sports	documentary_adrenaline	BBC_wildlife_eagle
# frames	246	246	101	246	195	107
fps	30	30	30	30	30	30

360 visual surprise maps $S_V(t; f, d)$. A window width for computing NCC depends on the length of an auditory event, namely the length $|T_{S,i}|$ of the segment. Through this process, 360 maps are obtained, each representing how synchronized every pixel in the corresponding visual surprise map is with the auditory surprise signal.

(3) Every synchronization map is finally averaged over pixels to obtain a sequence $c(t; f, d)$ that describes how synchronized the visual surprise map $S_V(t; f, d)$ is with the auditory surprise $S_A(t)$.

3.4 Features Selection

Once we have detected visual events synchronized with auditory events, the next step is to find dominant visual features in the detected events and to emphasize them to compute the final surprise map. This harmonizes our model with the finding that auditory attention modulates visual attention in a feature-specific manner [1].

First, the number of samples greater than a pre-specified threshold θ_c is counted in every sequence $c(t; f, d)$ to create a histogram with 360 (= the number of visual features) bins. Since every visual surprise map corresponds to a specific pair of feature type f and cascade level d (cf. Section 3.2), the histogram represents how dominant each feature is in synchronized audio-visual events. We do not accumulate $c(t; f, d)$ over time t to create a histogram because surprise values (accordingly, cross correlation values as well) at different frames cannot be compared in principle. We instead evaluate at each time t whether $c(t; f, d)$ is greater than a threshold or not, and if it is we vote for the corresponding visual feature.

Feature selection can be achieved by just binarizing the histogram, where a threshold for the binarization is adaptively chosen so that its slight change significantly impacts on the number of selected features. Only the visual surprise maps of the selected features (with active in the binarized histogram) are accumulated to form the final surprise map. In this way, our proposed model uses a smaller number of visual features than 360 for forming the final map.

4 Experiments

We experimentally evaluated our proposed model. We selected 6 video clips (advert bbc4 bees, advert bbc4 library, sports kendo, basketball of sports, documentary adrenaline, BBC wildlife eagle), all of which are provided by the DIEM project[1]. Table 1 illustrates the details of the video clips. We showed them to 15 human subjects. While

[1] `http://thediemproject.wordpress.com`

Table 2. NSS with optimal threshold values (bold letters: highest NSS for each video)

	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
baseline (Itti2009 [8])	0.299	1.524	0.636	**2.763**	1.275	0.450
proposed (θ_s, θ_c: optimal)	**0.935**	**1.842**	**0.801**	**2.763**	**1.287**	**0.621**
proposed ($\theta_s = 0, \theta_c$: optimal)	0.605	1.543	**0.801**	2.485	1.275	0.589

Table 3. Selected features using the optimal thresholds

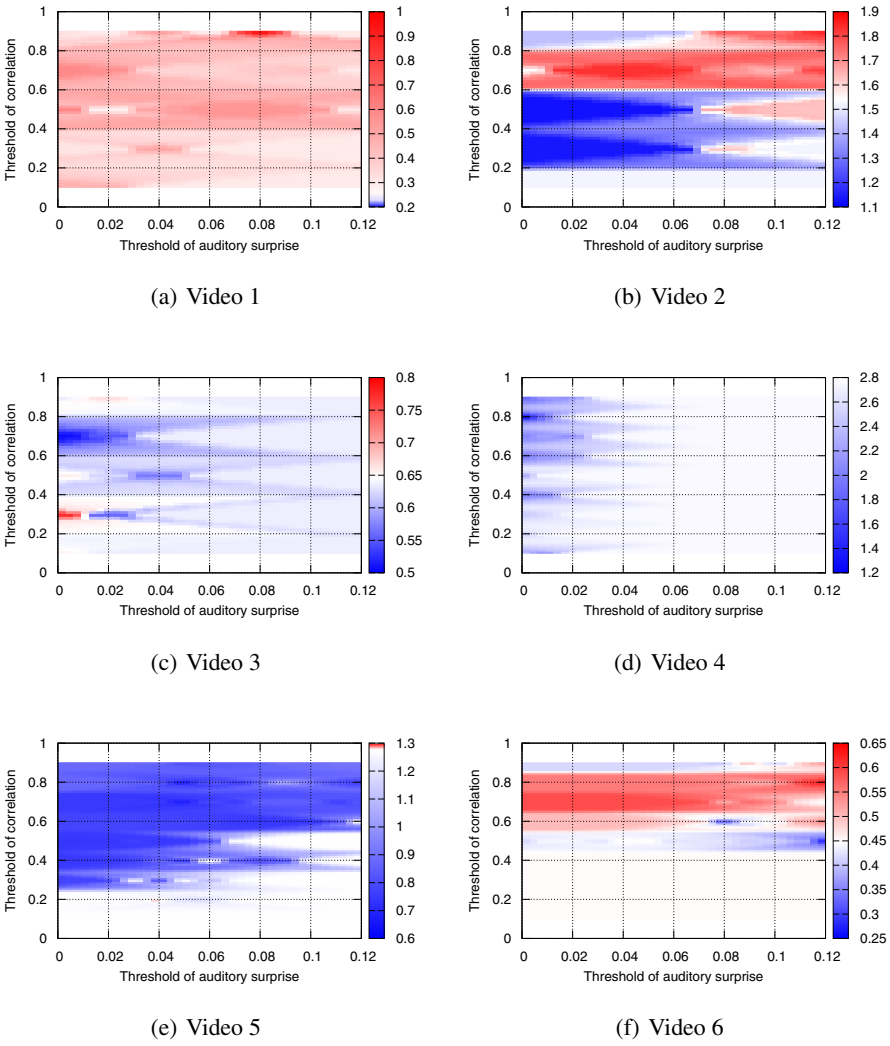
	Baseline	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
Intensity	30	0	8	8	30	0	6
Color	60	4	17	27	60	8	23
Orientation	120	0	46	39	120	0	7
Onset	30	0	0	0	30	1	0
Motion	120	0	0	0	120	14	0
Total	360	4	71	74	360	23	36

the subjects were watching the video clips, their eye movements were recorded using an eye tracker Tobii TX300. Note that we showed all the video clips to the subjects together with audio signals and originally collected their eye movements rather than directly exploiting the accompanying eye traces by the DIEM project. We extracted gaze points from the eye movements by removing micro-saccades. Namely, we removed all the eye movements greater than 2.12 pixels per millisecond, and identified all the remainings as gaze points (ground truths for the evaluation). As a metric to quantify how well a model predicts actual human eye movements, we used the normalized scan-path saliency (NSS) [8] calculated from the gaze points.

We first evaluated how two thresholds, i.e., auditory surprise threshold θ_s and correlation threshold θ_c, have impact on NSS. We changed the two threshold values independently and averaged NSS scores over frames for each video. Fig. 2 visualizes the averaged NSS scores for each video in terms of the heat map. In this visualization, red areas indicate threshold pairs with better performance than the Bayesian surprise model [8] (we call this the baseline model) and blue ones are the opposite. Table 2 shows the optimal values for θ_s and θ_c found in Fig. 2 where the optimal values mean the values that produce best NSS in our model. We remark that we also show the optimal value for θ_c under the condition that[2] $\theta_s = 0$.

From Fig. 2, we see that for Videos 1, 2 and 6, the proposed model is in most cases superior to or compatible with the baseline model and that for Videos 3 and 4 the proposed models is almost compatible with the baseline model for any threshold values. In contrast, for Video 5 our model could not achieve the compatible level with the baseline in most cases. Table 2 shows that our model with optimal threshold values produced higher NSS scores than the baseline model except for Video 4. NSS scores for Video 4 are fairly high themselves and, moreover, the ball is the only moving object and almost all image features are similar over frames in the video; these bring difficulty in selecting

[2] $\theta_s = 0$ is equivalent to computing correlation between audio and visual surprises for all the frames in the video.



(a) Video 1 (b) Video 2

(c) Video 3 (d) Video 4

(e) Video 5 (f) Video 6

Fig. 2. NSS scores under different threshold values (horizontal axis: auditory surprise threshold θ_s; vertical axis: correlation threshold θ_c). Red areas indicate thresholds with better performance than the baseline and blue ones are the opposite.

features synchronized with auditory events, which results in failure of further improvement of NSS scores. We also observe in Table 2 that we have better NSS scores by restricting the correlation computation only to the frames where auditory events occur than by correlation computation using all the frames in the video.

We see in Fig. 2 that for the same auditory surprise threshold value, a larger correlation threshold value tends to achieve better NSS for several videos while for the same

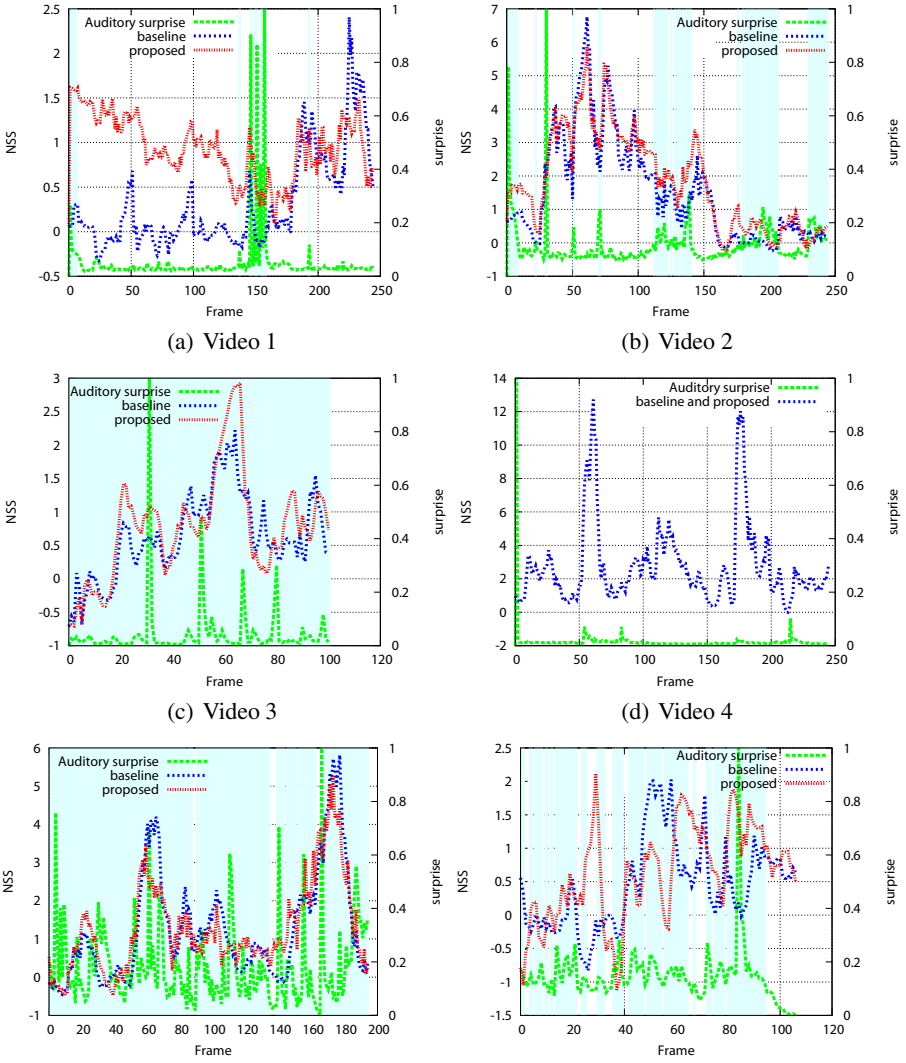


Fig. 3. NSS scores with the optimal thresholds and auditory surprises (horizontal axis: frame; vertical axis: NSS or normalized auditory surprise; light blue frame: surprising auditory events).

correlation threshold value, the auditory surprise threshold does not affect significant difference on NSS. This observation indicates that image features that have stronger correlation with the audio signal contribute to increase NSS more for any level of auditory surprise. This is consistent with the psychophysical finding that audio-visual temporal alignment leads to benefits in visual attention if changes in the component signals are both synchronized and transient [4]. We should note that two thresholds, in particular the correlation threshold, should be carefully chosen depending on the video for better performance.

To each video, we used the optimal threshold values shown in Table 2 to compute NSS scores for frames in the video. We illustrated in Fig. 3 detailed time trends of NSS scores for each video where the blue and red lines represent NSS scores for baseline and proposed models respectively. Note that in Video 4, the NSS score with the optimal threshold values was the same as that by the baseline model; blue and red lines are identical. Time trends of normalized auditory surprise are also shown in Fig. 3 using green lines. We see that in many frames of all the videos, the proposed model outperformed the baseline model. This is remarkable for Video 1. For Video 6, which model has a better NSS score heavily depends on the frame though the averaged NSS for our model is better. This is because Video 6 has several sudden changes in the scene and thus image features that are synchronized with auditory events also easily change.

Table 3 shows the number of selected visual features by the our model with the optimal threshold values. We can see that only a small fraction of 360 types of features were selected. We also observe that categories such as intensity or color of selected features highly depend on each input video. This is reasonable because what image features are closely correlated with auditory events depends on the video. This also justifies our implementation with a two-pass algorithm.

In order to show the effectiveness of our proposed model, we compared performances with the state-of-the-art models in addition to the baseline model [8]. They are the saliency map model [7], and the audio-visual attention model using the sound localization [14]. We also compared our model with the model, hereafter called the random feature selection model, in which we randomly selected a given number of image features among all image features used in [8], where the number of features to be selected was set in accordance with the number of image features determined by the optimal threshold values (see Table 2). We remark that NSS of the random feature selection model was computed as the average of NSS scores over 20 trials in random feature selections because selected image features vary at each iteration and produce a different NSS score.

Figure 4 illustrates averages of NSS scores over frames for each video and for each model. Note that the random feature selection model has error bars representing the standard deviation over 20 trials. We remark that all the image features (360 features) are selected for Video 4 and thus three models (the random feature selection model, the baseline model and our model) produce the same NSS score.

We see in Fig. 4 that our proposed model produced best NSS scores for all the video, outperforming the other models. The audio-visual attention model using the sound localization [14] did not achieve a good level in spite that it additionally uses auditory signals. This will be because it could not detect the sound source location accurately. Interestingly, the random feature selection model tends to outperform the baseline model. This indicates that using all the image features does not necessarily perform better. Using a smaller number of image features may be better. The number of selected features in our model may suggest such numbers though further investigation on required number of features is left for future work.

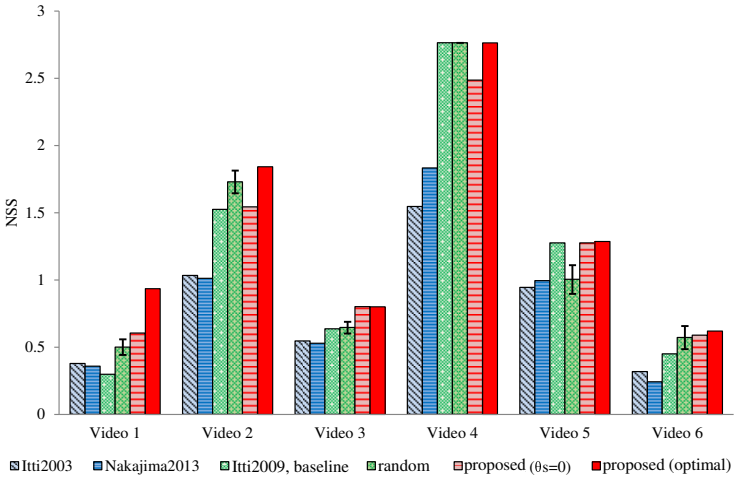


Fig. 4. Comparison of NSS averaged over frames for each video

5 Concluding Remarks

This paper proposed a novel computational model of human visual attention driven by auditory cues. Our model first detects synchronized and transient audio-visual events using a framework of Bayesian surprise and then selects dominant visual features in the detected events to form the final output. Our approach stands on using auditory features as a synchrony cue for selecting visual features. Differently from just fusing audio-visual information, our approach boosts the ability of visual information by selecting visual features synchronized with surprising auditory events. We remark that our approach is in line with recent psychophysical findings as well. The experimental evaluation with human eye movements demonstrated that our model outperformed the state-of-the-art models, in particular, the baseline model [8] in spite that we used a smaller number of visual features.

We used correlation to evaluate synchronization between audio and visual surprises. Mutual information can be also used as a measure for synchronization [17]. We are currently working for using mutual information instead of correlation. Our proposed model provides just one way to incorporate auditory cues into a computational model of human visual attention. We can thus improve our model into several directions in future, e.g. the introduction to adaptive image feature selection depending on the auditory event or the location in the image and machine learning strategies for capturing generic structures of audio-visual events.

References

1. Ahveninen, J., Jaaskelainen, I.P., Belliveau, J.W., Hamalainen, M., Lin, F.H., Raij, T.: Dissociable influences of auditory object *vs.* spatial attention on visual system oscillatory activity. PLoS One 7(6), e38511 (2012)
2. Begum, M., Karray, F.: Visual attention for robotic cognition: A survey. IEEE Transactions on Autonomous Mental Development 3(1), 92–105 (2011)
3. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 185–207 (2013)
4. Van der Burg, E., Cass, J., Olivers, C.N.L., Theeuwes, J., Alais, D.: Efficient visual search from synchronized auditory signals requires transient audiovisual events. PLoS One 5(5), e10664 (2010)
5. Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., Avrithis, Y.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia 15(7), 1553–1568 (2013)
6. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 989–1005 (2009)
7. Itti, L., Dhavale, N., Pighin, F.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Proc. SPIE 48th Annual International Symposium on Optical Science and Technology, vol. 5200, pp. 64–78. SPIE Press, Bellingham (2003)
8. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49(10), 1295–1306 (2009)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
10. Kayser, C., Petkov, C., Lippert, M., Logothesis, N.: Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology 15, 1943–1947 (2005)
11. Kimura, A., Yonetani, R., Hirayama, T.: Computational models of human visual attention and their implementations: A survey. IEICE Transactions 96-D(3), 562–578 (2013)
12. Ma, Y.F., Hua, X.S., Lu, L., Zhang, H.J.: A generic framework of user attention model and its application in video summarization. IEEE Transactions on Multimedia 7(5), 907–919 (2005)
13. Miyazato, K., Kimura, A., Takagi, S., Yamato, J.: Real-time estimation of human visual attention with dynamic Bayesian network and MCMC-based particle filter. In: ICME, pp. 250–257. IEEE (2009)
14. Nakajima, J., Sugimoto, A., Kawamoto, K.: Incorporating audio signals into constructing a visual saliency map. In: Klette, R., Rivera, M., Satoh, S. (eds.) PSIVT 2013. LNCS, vol. 8333, pp. 468–480. Springer, Heidelberg (2014)
15. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Video summarization and scene detection by graph modeling. IEEE Transactions on Circuits and Systems for Video Technology 15(2), 296–305 (2005)
16. Pang, D., Kimura, A., Takeuchi, T., Yamato, J., Kashino, K.: A stochastic model of selective visual attention with a dynamic Bayesian network. In: Proc. IEEE International Conference on Multimedia and Expo. (ICME), pp. 1073–1076. IEEE (2008)
17. Rolf, M., Asada, M.: Visual attention by audiovisual signal-level synchrony. In: Proc. 9th ACM/IEEE International Conference on Human-Robot Interaction Workshop on Attention Models in Robotics: Visual Systems for Better HRI (2014)
18. Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., Pfeifer, R.: Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 962–967 (2008)

19. Schauerte, B., Kühn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: Proc. 24th International Conference on Intelligent Robots and Systems (IROS). IEEE/RSJ (2011)
20. Schauerte, B., Stiefelhagen, R.: Wow! Bayesian surprise for salient acoustic event detection. In: Proc. 38th International Conference on Acoustics, Speech, and Signal Processing, (ICASSP) (2013)
21. Spexard, T., Hanheide, M., Sagerer, G.: Human-oriented interaction with an anthropomorphic robot. IEEE Transactions on Robotics 23(5), 852–862 (2007)
22. Tsuchida, T., Cottrell, G.: Auditory saliency using natural statistics. In: Proc. Annual Meeting of the Cognitive Science (CogSci), pp. 1048–1053 (2012)
23. Wolfe, J., Cave, K., Franzel, S.: Guided search: an alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance 15(3), 419–433 (1989)
24. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision 8(7) (2008)