Fixation Prediction through Multimodal Analysis

Xiongkuo Min, Guangtao Zhai, Chunjia Hu, Ke Gu

Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
{minxiongkuo,zhaiguangtao,huchunjia.sjtu,gukesjtuee}@gmail.com

***Abstract*—In this paper, we propose to predict human fixations by incorporating both audio and visual cues. Traditional visual attention models generally make the utmost of stimuli's visual features, while discarding all audio information. But in the real world, we human beings not only direct our gaze according to visual saliency but also may be attracted by some salient audio. Psychological experiments show that audio may have some influence on visual attention, and subjects tend to be attracted the sound sources. Therefore, we propose to fuse both audio and visual information to predict fixations. In our framework, we first localize the moving-sounding objects through multimodal analysis and generate an audio attention map, in which greater value denotes higher possibility of a position being the sound source. Then we calculate the spatial and temporal attention maps using only the visual modality. At last, the audio, spatial and temporal attention maps are fused, generating our final audio-visual saliency map. We gather a set of videos and collect eye-tracking data under audio-visual test conditions. Experiment results show that we can achieve better performance when considering both audio and visual cues.**

***Index Terms*—Audio-visual attention, multimodal analysis, saliency, fixation prediction, attention fusion**

I. Introduction

Visual attention has long been an important research topic in areas of psychology, image processing and computer vision. During recent years, many visual attention computational models aiming at predicting eye fixation, detecting salient object and generating objectness proposal have been proposed [1], [2]. Most models utilize low-level visual features such as intensity, color and orientation [3] to highlight positions which are distinctly different from its surroundings. Some models also take some high-level cognitive features into account, for example face and person detectors [4]. For dynamic scenes, many models also consider motion features [5], [6]. In spite of various kinds of features used to model visual attention, all these features are visual features. An important thing is that almost all visual attention models leave audio information behind. Existing visual attention databases are mostly built under the visual test condition that subjects hear no audio. Whereas, some psychological works have shown that audio does have some impact on visual attention [7], [8], [9]. Antoine et al. [7] verified that sound impacted on eye movements, and the impact varied across time. Song et al. [8] found that only certain kinds of sound especially human voice had an influence on eye movements. Min et al. [9] demonstrated that the impact of audio was up to its consistency with visual signals. All these psychological works are based on subjective eye-tracking experiments.

Although plenty of psychological works have verified the influence of audio on visual attention, little efforts have been devoted to applying those findings to visual attention modeling. Several researchers have made an attempt in constructing an audio-visual attention model [10], [11], [12]. Chen et al. [10] captured eye-tracking data for a set of image-audio pairs. Experiments shown that coherent audio information helps to enhance the saliency of corresponding visual target. A framework was also proposed to predict fixation in scenes of image-viewing with the influence of different audio. Conversation scenes were investigated in [11], and they proposed an audio-visual saliency model for natural conversation scenes based on the fact that the speaking faces were generally much more salient compared with others. Treating the sound sources as the most visually salient parts of videos, Lee et al. [12] presented a foveated coding method using audio-visual focus of attention.

In this work, we concentrate on constructing audio-visual attention model to predict fixation in videos. Different from aforementioned works which only work in specific scenes such as image-audio pairs [10] or conversation scenes [11], our framework is applicable to more practical and general conditions. Based on the finding that audio affects visual attention and the sound sources are strong cues for visual attention [7], [8], [9], we try to model visual attention from both visual and audio perspective. A framework of our approach is illustrated in Fig. 1. The spatial and temporal visual attention maps are calculated directly from the video stream, like traditional saliency models. Whereas for the audio, we attempt to localize the moving-sounding objects through mutimodal analysis. The localization result is taken as our audio attention map. At last, the audio and visual attention maps are fused to the final audio-visual saliency map.

The remainder of this paper is organized as follows. Section II detailedly describes how to model audio-visual attention. In Section III, we present the subjective eye-tracking experiment. The effectiveness of the presented audio-visual attention model is also verified in this section. Section IV concludes this paper.

II. Audio-visual attention model

Following the framework illustrated in Fig. 1, we first model audio and visual attention respectively, then audio and visual attention maps are fused to the final audio-visual saliency map.

IEEE VCIP'15, Dec. 13 - Dec. 16, 2015, Singapore.
978-1-4673-7314-2/15/$31.00

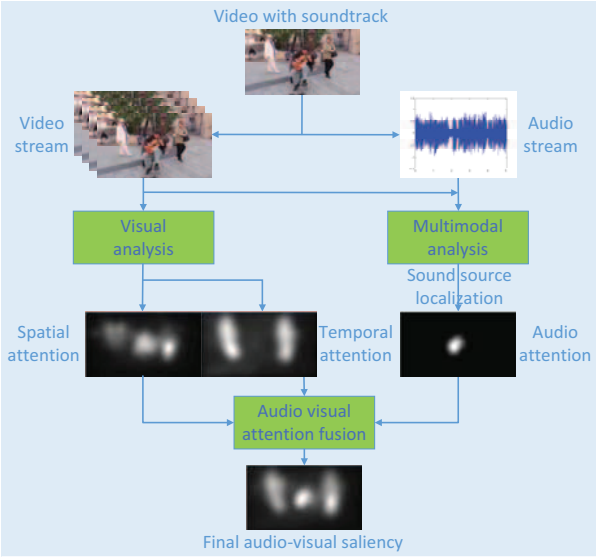

Fig. 1. Framework of the audio-visual attention model.

A. Audio Attention Model

We try to localize the moving-sounding objects in videos and take the localization result as the audio attention map. For video captured using a single microphone, it is possible to localize the moving-sounding objects through multimodal correlation analysis since the objects' motion pattern and the audio variation pattern are highly correlated in such scenes [13], [14]. We adopt a correlation-after-segmentation approach similar to [13], but with some modifications to make it more appropriate for our purpose. Fig. 2 illustrates the flow diagram of the localization method used in this paper. We first segment the entire video into a specific number of appearance-motion-coherent spatial-temporal regions (STRs). For each STR, velocity and acceleration derived from optical flow are used as visual features. Mel-frequency Cepstral Coefficients (MFCCs) and its first order derivatives (MFCC_Ds) are calculated to represent the audio. Finally, canonical correlation analysis (CCA) is utilized to determine the STRs whose visual features are the most correlated with audio features.

1) Visual analysis: The purpose of visual analysis is to segment the video into K supervoxels $SV_k(t)(t = 1, ..., T; k = 1, ..., K)$ and represent each supervoxel with some motion features. We use a graph-based streaming hierarchical method [15] to segment the video. Motion features are velocity and acceleration derived from optical flows [16]:

$$\mathbf{vel} = \mathbf{U}^{+}(i, j, t) \tag{1}$$

$$\mathbf{acl} = \mathbf{U}^{+}(i, j, t) - (-\mathbf{U}^{-}(i, j, t)) \tag{2}$$

where $\mathbf{U}^{+}(i, j, t)$ represents forward optical flow from frame F_t to F_{t+1}, $\mathbf{U}^{-}(i, j, t)$ represents backward optical flow from frame F_t to F_{t-1}, (i, j) denotes pixel position. Then $SV_k(t)$ can be descried by the mean velocity and acceleration magnitude of pixels belong to $SV_k(t)$. We select the most dominate m_1 supervoxels for velocity and m_2 supervoxels for acceleration according to their variances along time axis. Finally we can use matrix $\mathbf{v} = [\mathbf{v}_1, ..., \mathbf{v}_T]$ to characterize the whole video, where $\mathbf{v}_t, t = 1, ..., T$ is a $M = m_1 + m_2$ dimension vector denotes the visual features for frame F_t.

2) Audio analysis: We assume that the audio signal is dominated by the sound emitting from the target moving-sounding objects. We extract $N/2$ MFCCs and its first order derivatives as audio feature. Then audio signal can be characterized by $\mathbf{a} = [\mathbf{a}_1, ..., \mathbf{a}_T]$, where $\mathbf{a}_t, t = 1, ..., T$ is a N dimension vector used to feature the t^{th} windowed audio signal.

3) Visual-audio correlation: The goal of visual-audio correlation is to detect the supervoxels, i.e., the dimensions of $\mathbf{v}$, that maximize its correlation with audio $\mathbf{a}$. Common correlation methods may suffer from the problem that the video and audio signal are described in distinctively different fields. Canonical Correlation Analysis (CCA) is a classic yet efficient method that can perform correlation analysis after project signals of different modality to a common coordinate system. Precisely in our work, CCA seeks pairs of canonical bases $\mathbf{w}_v$ and $\mathbf{w}_a$ that maximize the correlation between projections $\mathbf{w}_v^T\mathbf{v}$ and $\mathbf{w}_a^T\mathbf{a}$ [17]:

$$(\mathbf{w}_v, \mathbf{w}_a) = \arg\max_{\mathbf{w}_v, \mathbf{w}_a} \mathtt{Corr}(\mathbf{w}_v^T\mathbf{v}, \mathbf{w}_a^T\mathbf{a}) \tag{3}$$

Eq. (3) has a closed form solution as an eigenvalue problem:

$$\begin{cases} C_{vv}^{-1}C_{va}C_{aa}^{-1}C_{av}\mathbf{w}_v = \rho^2\mathbf{w}_v \\ C_{aa}^{-1}C_{av}C_{vv}^{-1}C_{va}\mathbf{w}_a = \rho^2\mathbf{w}_a \end{cases} \tag{4}$$

where C_{vv} and C_{aa} denote the covariance matrices of $\mathbf{v}$ and $\mathbf{a}$ respectively, C_{va} is the cross-covariance matrix of the vectors $\mathbf{v}$ and $\mathbf{a}$. Solving Eq. (3) is equivalent to finding the largest eigenvalue ρ_1^2 and corresponding eigenvectors $\mathbf{w}_{v^1}, \mathbf{w}_{a^1}$ in Eq. (4). Larger ρ_1^2 denotes better correlation between video and audio. Moreover, in $\mathbf{w}_{v^1}$, the components with higher magnitude values contribute more to the maximum correlation.

We generate a correlation map according to $\mathbf{w}_{v^1}$. Normalised components of $\mathbf{w}_{v^1}$ lager than a threshold and corresponding supervoxels are selected as candidates. In the correlation map, values of all pixels belonging to each candidate supervoxel are set to corresponding normalized $\mathbf{w}_{v^1}$ component value, while others are set to 0. The correlation

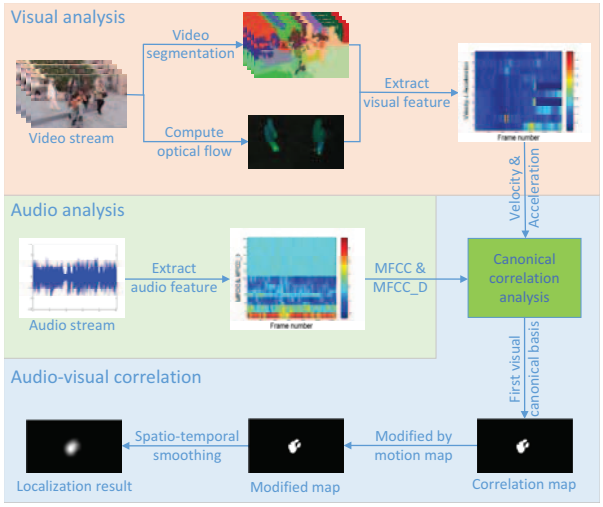

Fig. 2. Flow diagram of the moving-sounding objects localization method.

map is further modified by multiplying a motion map to remove tiny motions. Where in our motion map, values of pixels belonging to each supervoxel are set to this supervoxel's velocity variance along the time axis. Finally the modified map is spatial-temporally smoothed to the final localization result, and it is taken as the audio attention map S_a.

B. Visual Attention Model

Similar to conventional video saliency models, we predict visual attention from both spatial and temporal aspects.

1) Spatial attention model: In recent years, dozens of saliency models have achieved good performance for static images. Since the main purpose of this work is to demonstrate the superiority of visual-audio attention fusion, we choose 8 typical image saliency algorithms to model spatial attention: IT [3], GBVS [18], SUN [19], SR [20], PFT [5], SeR [6], Judd [4] and RC [21]. The spatial attention map is denoted as S_s in following sections.

2) Temporal attention model: Object motion is an important cue for visual attention. Optical flow is often used to describe the local motion of videos [16]. We adopt optical flow based motion estimation for temporal attention to reduce computation since we have calculated optical flow for each frame in the audio attention model. The temporal attention map S_t can be calculated by:

$$S_t = g * ||\mathbf{vel}|| \tag{5}$$

where g is a Gaussian kernel and **vel** is the velocity acquired by Eq. (1).

C. Audio-Visual Attention Fusion

The final stage is to generate the final audio-visual attention map by fusing the audio and visual attention maps:

$$S = f(S_s, S_t, S_a) \tag{6}$$

where S is the final audio-visual saliency map, S_s, S_t, S_a are spatial, temporal and audio attention maps respectively, f is the fusion function. Because of its simplicity and generalizability, we choose normalization and summation as our fusion method. Then f can be described as:

$$f : S_i \rightarrow \mathcal{N}\Big(\sum_i \mathcal{N}(S_i)\Big), i \in \{s, t, a\} \tag{7}$$

where $\mathcal{N}$ is a normalization operator to normalize all attention maps to the same dynamic range, i.e. $[0, 1]$.

III. Experiments and Results

A. Subjective Eye-tracking Experiments

Since there is no public audio-visual attention database available for our purpose, we perform eye-tracking experiments on 30 test videos. Videos are gathered from [13], [14] and YouTube. Collected videos contain a variety of scenes, e.g. playing musical instruments, playing or kicking the ball, speaking face and conversations. The lengths of videos range from 5 to 10 seconds. We use Tobii T120 Eye Tracker to collect eye movement data with a free-viewing task condition. All videos contain soundtracks, and eye-tracking experiments are conducted in an audio-visual condition which subjects watch the video and listen to the soundtrack synchronously. The tracking distance is around 60 cm, and the sampling rate is set to 120 Hz. Tobii T120 has a screen resolution of 1280×720 pixels. During the tests, videos are linearly rescaled to fit the maximum resolution of the screen, but we do not change the aspect ratio of videos. A total of 16 college students participate in our experiments. But data from 2 subjects are abandoned because of tracking problems.

TABLE I
Model Parameters

Category	Parameter	Value
Video segmentation	Merging threshold (pixel level)	5
	Merging threshold (hierarchical level)	200
	Minimum segment size	100
	No. of frames in a clip	15
	No. of supervoxels	About 25
Optical flow	Regularization weight	0.012
	Downsample ratio	0.5
	Width of the coarsest level	40
	No. of fixed point iterations	7 (outer) 1 (inner)
	No. of SOR iterations	30
Visual feature	No. of top supervoxels	5 (velocity) 5 (acceleration)
Audio feature	No. of MFCCs	10
	No. of MFCC_Ds	10
AV correlation	Threshold	0.4
	SD. of the Gaussian kernel	10 (spatial) 5 (temporal)
Temporal attention	SD. of the Gaussian kernel	10

B. Model implementation details

In the implementation, all videos are analyzed at its original frame rate, but we down-sample the videos to make them have a maximum width or height of 240 pixels without change of videos' aspect ratio. Before audio processing, the audio signal is framed to have the same number of frames as video, and the framing windows are 50% overlapped. The adopted video segmentation approach [15] is a hierarchical method. We choose the desired level such that the final number of supervoxels is most close to 25. More model implementation parameters can be found in Table I. Note that parameters are mainly from third party algorithms [15], [16].

C. Experiment Results

Similar to study [22], we use 3 saliency evaluation metrics: AUC (Area under the Receiver Operating Characteristic curve), Linear Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS). Based on the gathered videos and corresponding eye-movement data, we demonstrate the effectiveness of the proposed framework by evaluation of three types of attention maps: S_s, $f(S_s, S_t)$ and $f(S_s, S_t, S_a)$. As mentioned, S_s denotes spatial saliency map. In this paper, 8 typical image saliency algorithms are measured. $f(S_s, S_t)$ represents the combination of spatial and temporal saliency maps and $f(S_s, S_t, S_a)$ stands for the fusion of spatial, temporal and audio attention maps. f is the fusion method described in Section II-C. Experiment results are listed in Table II, where

TABLE II
PERFORMANCE EVALUATION OF 8 IMAGE SALIENCY ALGORITHMS AND ITS COMBINATION WITH TEMPORAL AND AUDIO ATTENTION MAPS. S: SPATIAL, ST: SPATIAL-TEMPORAL, STA: SPATIAL-TEMPORAL-AUDIO.

Metrics	AUC			CC			NSS		
	S	ST	STA	S	ST	STA	S	ST	STA
IT	0.834	0.873	**0.895**	0.314	0.386	**0.448**	1.343	1.703	**2.028**
GBVS	0.865	0.887	**0.901**	0.358	0.410	**0.466**	1.582	1.836	**2.125**
SUN	0.703	0.832	**0.878**	0.167	0.315	**0.412**	0.732	1.424	**1.895**
SR	0.709	0.839	**0.875**	0.233	0.345	**0.432**	1.036	1.553	**1.981**
PFT	0.733	0.833	**0.876**	0.202	0.340	**0.431**	0.904	1.526	**1.973**
SeR	0.727	0.821	**0.873**	0.215	0.319	**0.406**	0.937	1.414	**1.843**
Judd	0.857	0.875	**0.892**	0.324	0.396	**0.456**	1.448	1.790	**2.098**
RC	0.742	0.837	**0.880**	0.202	0.333	**0.420**	0.854	1.483	**1.922**

"S", "ST" and "STA" denote the performance of S_s, $f(S_s, S_t)$ and $f(S_s, S_t, S_a)$ respectively. Results listed in this table are average performance of all test videos. For most tested saliency models, $f(S_s, S_t)$ performs better than S_s not surprisingly, since motion is an important incentive for visual attention. But $f(S_s, S_t, S_a)$ performs even better than $f(S_s, S_t)$, which is a quantitative verification of our framework. That is, audio has some influence on visual attention and we can promote visual attention prediction by incorporating the audio cues.

Fig. 3 is an intuitive illustration of related saliency maps. It is a typical example that audio information matters. In this example, the left talking face is a strong attractor to visual attention, but it is not easy to be detected by purely visual analysis if no other high-level cognitive factors are considered. Through multimodal analysis, we can locate the talking face, thus $f(S_s, S_t, S_a)$ works better than S_s and $f(S_s, S_t)$.

IV. CONCLUSION

Audio information is an indispensable part of multimedia content, but it is rarely considered in visual attention models. Psychological findings show that the sound source is a strong incentive for visual attention. We apply it to fixation prediction. Through audio-visual correlation analysis, we locate the moving-sounding objects and generate an audio attention map for each frame. The audio attention maps are further fused with conventional visual attention maps. The efficiency of generated audio-visual saliency maps is verified with gathered videos and eye-movement data. Experiment results also show that incorporating audio information may not always aid visual attention prediction because of some factors such as the accuracy of moving-sounding objects localization. Comprehensive investigation of audio information's helpfulness in visual attention modeling and more robust audio-visual saliency models are goals of our future work.

ACKNOWLEDGMENT

This work was supported in part by NSFC (61025005, 61371146, 61221001), 973 Program (2010CB731401) and FANEDD (201339).

REFERENCES

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

[2] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *ArXiv preprint*, 2014.

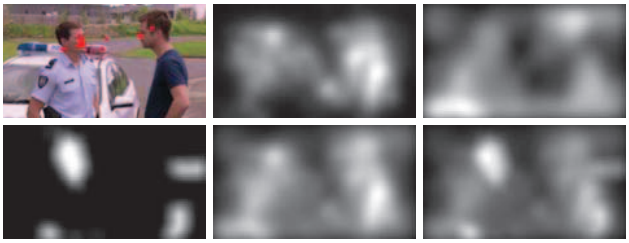

Fig. 3. An example of related saliency maps. From left to right, from top to bottom: frame image with gaze points, S_s, S_t, S_a, $f(S_s, S_t)$ and $f(S_s, S_t, S_a)$. S_s is calculated from Itti's method [3] in this figure.

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

[4] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *12th IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[5] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, 2009.

[7] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[8] G. Song, D. Pellerin, and L. Granjon, "Different types of sounds influence gaze differently in videos," *Journal of Eye Movement Research*, vol. 6, no. 4, pp. 1–13, 2013.

[9] X. Min, G. Zhai, Z. Gao, C. Hu, and X. Yang, "Sound influences visual attention discriminately in videos," in *6th IEEE International Workshop on Quality of Multimedia Experience*, Sept 2014, pp. 153–158.

[10] Y. Chen, T. Nguyen, M. Kankanhalli, J. Yuan, S. Yan, and M. Wang, "Audio matters in visual attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 11, pp. 1992–2003, Nov 2014.

[11] A. Coutrot and N. Guyader, "An audiovisual attention model for natural conversation scenes," in *IEEE International Conference on Image Processing*, Oct 2014, pp. 1100–1104.

[12] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1322–1331, 2011.

[13] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, 2013.

[14] K. Li, J. Ye, and K. A. Hua, "What's making that sound?" in *ACM International Conference on Multimedia*, 2014, pp. 147–156.

[15] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *European Conference on Computer Vision*. Springer, 2012, pp. 626–639.

[16] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Citeseer, 2009.

[17] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.

[19] 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, p. 32, 2008.

[20] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[21] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 409–416.

[22] A. Borji, D. 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, Jan 2013.