Fixation Prediction through Multimodal Analysis

XIONGKUO MIN, GUANGTAO ZHAI, KE GU, and XIAOKANG YANG,
Shanghai Jiao Tong University

In this article, we propose to predict human eye fixation through incorporating both audio and visual cues. Traditional visual attention models generally make the utmost of stimuli's visual features, yet they bypass all audio information. In the real world, however, we not only direct our gaze according to visual saliency, but also are attracted by salient audio cues. Psychological experiments show that audio has an influence on visual attention, and subjects tend to be attracted by the sound sources. Therefore, we propose fusing both audio and visual information to predict eye fixation. In our proposed framework, we first localize the moving–sound-generating objects through multimodal analysis and generate an audio attention map. Then, we calculate the spatial and temporal attention maps using the visual modality. Finally, the audio, spatial, and temporal attention maps are fused to generate the final audiovisual saliency map. The proposed method is applicable to scenes containing moving–sound-generating objects. We gather a set of video sequences and collect eye-tracking data under an audiovisual test condition. Experiment results show that we can achieve better eye fixation prediction performance when taking both audio and visual cues into consideration, especially in some typical scenes in which object motion and audio are highly correlated.

CCS Concepts: • **Computing methodologies** → **Interest point and salient region detections**; *Scene understanding*; *Vision for robotics*; • **Information systems** → *Spatial-temporal systems;*

Additional Key Words and Phrases: Audiovisual attention, multimodal analysis, saliency, eye fixation prediction, attention fusion

ACM Reference Format:
Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. 2016. Fixation prediction through multimodal analysis. ACM Trans. Multimedia Comput. Commun. Appl. 13, 1, Article 6 (October 2016), 23 pages.
DOI: http://dx.doi.org/10.1145/2996463

1. INTRODUCTION

Visual attention has long served as an important research topic in areas of psychology, image processing, and computer vision. Predicting where humans look can be of great use in numerous applications, such as image-quality assessment [Ninassi et al. 2007], video coding [Itti 2004], and automatic contrast enhancement [Gu et al. 2015b]. In recent years, many visual attention computational models have been developed [Borji and Itti 2013; Borji et al. 2014]. Most models utilize low-level visual features, such as intensity, color, and orientation [Itti et al. 1998; Harel et al. 2006], to highlight positions that are distinctly different from their surroundings. Some models also take

This work was supported in part by the National Natural Science Foundation of China under Grants 61422112, 61371146, 61521062, and 61527804; and the National High-Tech R&D Program of China under Grant 2015AA015905.
Authors' addresses: X. Min, G. Zhai, K. Gu, and X. Yang, Institute of Image Communication and Network Engineering, Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiao Tong University, Shanghai 200240, China; emails: minxiongkuo@gmail.com, zhaiguangtao@sjtu.edu.cn, guke.doctor@gmail.com, xkyang@sjtu.edu.cn.

DOI: http://dx.doi.org/10.1145/2996463

some high-level cognitive visual features into account, for example, using face [Cerf et al. 2008], text [Cerf et al. 2009], and person detectors [Judd et al. 2009]. For dynamic scenes, many models also consider motion features [Guo et al. 2008; Seo and Milanfar 2009; Kim et al. 2015].

In spite of various kinds of features used to model visual attention, almost all of these features used are visual based, and almost all visual-attention models leave audio information aside. Existing visual-attention databases are mostly built under the visual test condition in which subjects hear no audio, whereas some psychological works have shown that audio does have some impact on visual attention [Perrott et al. 1990; Vroomen and Gelder 2000; Coutrot and Guyader 2013; Song et al. 2013; Min et al. 2014]. In an early study, Perrott et al. [1990] showed that sound can guide attention to a visual target when auditory and visual signals came from the same position. Vroomen and Gelder [2000] demonstrated that concurrent auditory stimuli can enhance visual perception. Bao and Roy Choudhury [2010] used audio and sensory cues to identify temporal segment boundaries of important events in social videos.

In addition to traditional psychological experiments, the influence of sound was further verified through eye-tracking experiments [Coutrot and Guyader 2013; Song et al. 2013; Min et al. 2014]. Coutrot and Guyader [2013] investigated how sound impacted eye movements by controlling contents of the scenes. They found that sound influenced eye movements significantly only in videos containing faces and several moving objects. Song et al. [2013] found that the effect of sound was different depending on the sound types. They classified sound into various kinds, such as human voice, action, and music. Particular types of sound guided human eyes to the sound source, and the human voice showed the greatest impact on visual attention. In a preliminary work [Min et al. 2014], we demonstrated that the impact of audio was up to its consistency with visual signals. If the sound source was not the most visually salient object, subjects would be attracted by the sound sources to a certain degree. For example, in conversation scenes, subjects tended to focus more on the speaking person.

Although plenty of psychological works have verified the influence of audio on visual attention, few efforts have been devoted to applying those findings to visual attention modeling. Thus, in this work, we concentrate on constructing audiovisual attention models to predict eye fixation in video sequences. Based on the finding that audio affects visual attention in some circumstances and that sound sources are strong cues for visual attention in such scenes [Coutrot and Guyader 2013; Song et al. 2013; Min et al. 2014], we try to model visual attention from both audio and visual perspectives. A framework of our approach is illustrated in Figure 1. Like traditional saliency models, the spatial and temporal visual attention maps are calculated directly from the video stream. Spatial attention maps are calculated from single video frames using state-of-the-art image saliency algorithms. As demonstrated in Yantis and Jonides [1990], object motion can attract visual attention. In this work, we compute optical flows from adjacent frames to describe video sequences' motion characteristics, and the magnitudes of forward optical flows are taken as the temporal attention maps. For the audio, we attempt to localize the moving–sound-generating objects through multimodal analysis [Izadinia et al. 2013]. The localization result is taken as our audio attention map. Finally, the audio and visual attention maps are fused as the final audiovisual saliency map. The proposed method is applicable to scenes containing moving–sound-generating objects. In such scenes, it is feasible to localize and highlight the moving–sound-generating objects.

With the help of the sound source localization method [Izadinia et al. 2013], we can predict eye fixations better in scenes containing moving–sound-generating objects. A typical scene is that there are several moving objects, but only one is generating the sound. In such scenes, the influence of the sound is obvious, and it is also possible to

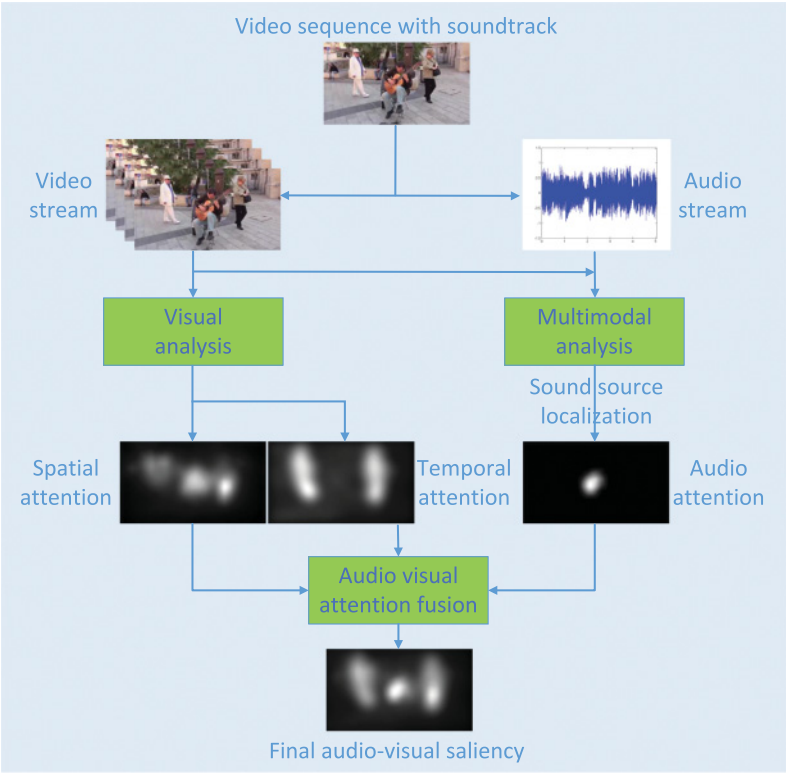

Fig. 1. Framework of the audio-visual attention model. The spatial and temporal visual attention maps are calculated directly from the video stream. The audio attention map are derived from the results of a sound source localization method.

localize the moving–sound-generating object. Scenes of this kind are also very common in our daily life. Note that the adopted and other state-of-the-art sound source localization approaches [Barzelay and Schechner 2007; Kidron et al. 2007; Izadinia et al. 2013; Li et al. 2014] can only work in scenes in which the object motion and sound are highly correlated, that is, the sound is generated by the object motion, whereas in other scenes in which the object motion and the sound are weakly correlated or uncorrelated, the localization methods are noneffective. Thus, our multimodal saliency models are also not effective in such scenes, for example, in videos containing background music or working loudspeakers.

The remainder of this article is organized as follows. In Section 2, we briefly review the related work. Section 3 describes in detail how to model audiovisual attention. In Section 4, we present the eye-tracking experiment. The effectiveness of the presented audiovisual attention model is also verified in this section. We present our conclusions in Section 5.

2. RELATED WORK

In this section, we briefly review some related works. First, we discuss the moving–sound-generating object localization methods. Then, we briefly discuss visual attention models. Finally, the important visual models that incorporate audio cues are reviewed.

2.1. Moving–sound-generating Object Localization

The key problem of audio attention modeling is to localize the moving–sound-generating objects in a video sequence. In a three-dimensional space, we can localize the sound source according to the small time discrimination of the sound's arrival at our ears [Jeffress 1948]. Minotto et al. [2014] performed voice-activity detection and sound-source localization for simultaneous speaker scenarios, using a camera and a microphone array. For video sequences captured using a single microphone, however, it is much more difficult to localize the sound source. Some researchers made attempts in such circumstances [Barzelay and Schechner 2007; Kidron et al. 2007; Izadinia et al. 2013; Li et al. 2014]. These methods were based on the premise that multimodal data originating from one source were highly correlated. Thus, auditory and visual data of the same video sequence could be used to localize visual events that generated the sound. Methods in Barzelay and Schechner [2007] and Kidron et al. [2007] were pixel-level localization. In Barzelay and Schechner [2007], audio and visual onsets were used to characterize each modality. Then, a coincidence-based measure was used to associate the visual and audio events. Kidron et al. [2007] employed a Canonical Correlation Analysis (CCA) method to perform cross-modal analysis.

To overcome the limitations of pixel-level analysis, some methods [Izadinia et al. 2013; Li et al. 2014] adopted object-level localization, in which correlation-after-segmentation methods were used. The authors first segmented the entire video sequence into a number of spatial-temporal regions (STRs). Then, visual features were extracted to represent each STR. The localization problem was reduced to detecting the STRs whose visual features were most correlated to the video sequence's audio features. We apply a localization approach similar to Izadinia et al. [2013], but with some modifications to make it more appropriate for our purpose of generating audio-attention maps. The original video segmentation process is replaced with a state-of-the-art video segmentation approach, and the final localized results are further modified. The localization results that highlight the sound sources are treated as audio-attention maps.

2.2. Visual-Attention Model

After decades of development, dozens of visual-attention models are available now. According to some recent review articles [Borji and Itti 2013; Borji et al. 2014], the research of visual-attention modelling can be classified into 3 closely related areas: fixation prediction, salient-object detection and object-proposal generation. Most early studies aim at fixation prediction [Itti et al. 1998; Harel et al. 2006; Zhang et al. 2008; Hou and Zhang 2007, 2009; Guo et al. 2008; Cerf et al. 2008; Seo and Milanfar 2009; Judd et al. 2009]. Models of this kind are developed to predict the positions of the human gaze. Eye-tracking data collected under a free-viewing condition is used to evaluate such models. Later, driven by saliency-based applications, salient-object detection models were proposed [Zhai and Shah 2006; Achanta et al. 2009; Cheng et al. 2011]. Models of this kind try to detect the most salient object as a whole. The generated saliency maps are used to segment the salient objects. More recently, some models have been introduced to generate object proposal [Cheng et al. 2014; Alexe et al. 2012]. The purpose is to generate a group of object regions to cover all objects in the scene. Human labeled data are used as the ground truth for the later two kinds of models. In this article, we mainly focus on the fixation prediction model, which is the most widely investigated and used type in the research community.

2.3. Visual-Attention Model Incorporating Audio Cues

As previously discussed in the introduction section, fixation-prediction models generally make the most of various low- or high-level visual features to highlight the most

distinctive positions. The extracted features are mostly visual based. Concerning the impact of audio, plenty of psychological studies have verified the influence of audio on visual attention, but few efforts have been devoted to applying those findings to visual-attention modeling. Several researchers have made attempts at constructing audiovisual-attention models [Ma et al. 2005; Evangelopoulos et al. 2013; Chen et al. 2014; Lee et al. 2011; Coutrot and Guyader 2014].

Ma et al. [2005] propose a user attention model that combines multiple sensory perceptions such as visual, audio stimulus, and some semantic understanding. The proposed model was successfully used in video summarization, which extracts important video content. Similarly, Evangelopoulos et al. [2013] fuse aural, visual, and textual attention for movie summarization. The goal of this attention model is to generate an attention curve along the time axis. High value indicates that the video content in the current moment is important. Audio is a rather important type of feature in these kinds of attention models. It is out the scope of this article since we are interested in predicting eye movement.

Chen et al. [2014] captured eye-tracking data for a set of image–audio pairs. Experiments showed that coherent audio information helped to enhance the saliency of the corresponding visual target. A framework was also proposed to predict eye fixation in scenes of image viewing with the influence of different audio. Their work is based on audio classification and visual object detection, which relies on training, and only a specific number of audio and object types were trained. Lee et al. [2011] present a foveated video coding method using audiovisual focus of attention. They localized and treated the sound source region as the most visually salient part of the video sequence. Then, the foveated video coding method assigned different quality levels to video frames according to the distance from a pixel to the localized sound source. In this method, they treated the sound source as the most salient location of the video. This method is not comprehensive since audio attention often integrates or competes with visual attention in the realistic scenes. We should also consider visual attention. In addition to general scenes, conversation scenes are specifically investigated in Coutrot and Guyader [2014]. They propose an audiovisual saliency model for natural conversation scenes based on the fact that the speaking faces were generally much more salient compared with others. Through increasing the saliency of the speakers, the proposed audiovisual saliency model performed better than visual attention models in which all faces were fairly treated.

In contrast to the aforementioned works, which are devoted only to specific scenes, such as image–audio pairs [Chen et al. 2014] or conversation scenes [Coutrot and Guyader 2014], our framework is less restrictive to video content. Compared with Lee et al. [2011], our method is more reasonable since we take both audio and visual attention into account, whereas Lee et al. [2011] treat only the sound source as the most salient part of the scene.

3. AUDIOVISUAL ATTENTION MODELING

Following the framework illustrated in Figure 1, we first model audio and visual attention, then audio and visual attention maps are fused to generate the final audiovisual saliency map. Details of the audiovisual attention modelling process are discussed later.

3.1. Audio Attention Modeling

As discussed in Section 1, we try to localize the moving-sound-generating objects in video sequences, and the localization result acts as the audio attention map. In scenes in which the audio is generated from object motion, we can localize the sound source since the object's motion pattern and the audio variation pattern are highly correlated

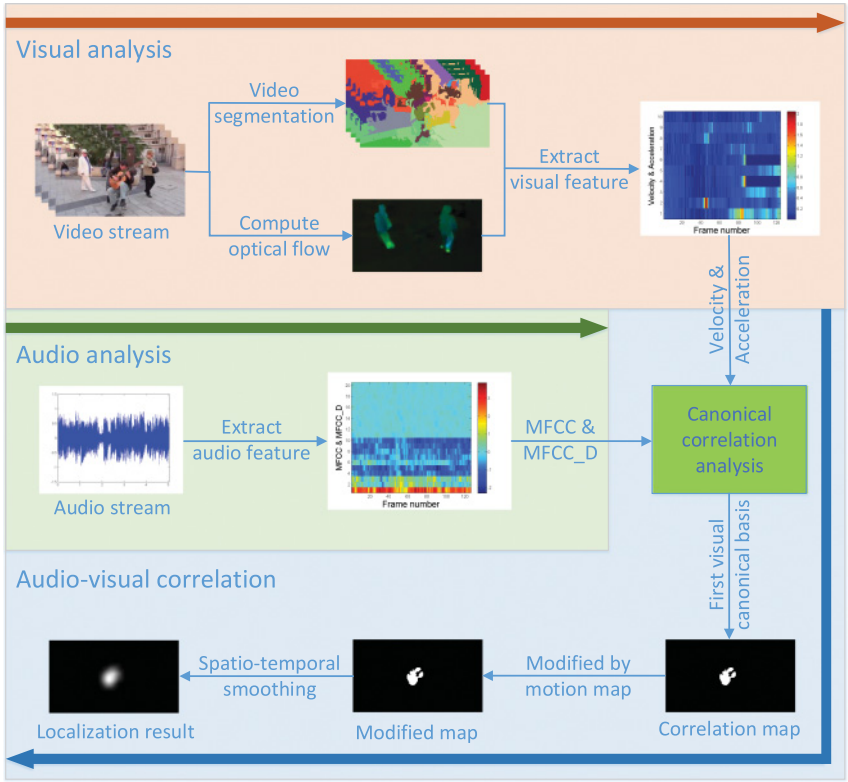

Fig. 2. Flow diagram of the moving-sound-generating objects localization method. The whole process can be divided into 3 parts: visual analysis, audio analysis, and audiovisual correlation.

in such scenes [Izadinia et al. 2013; Li et al. 2014; Kidron et al. 2005]. We adopt the same correlation-after-segmentation framework as Izadinia et al. [2013]. Figure 2 illustrates the flow diagram of the localization method used in this article. We first segment the entire video sequence into a specific number of appearance-motion-coherent STRs. For each STR, velocity and acceleration derived from optical flow are used as visual features. Mel-frequency Cepstral Coefficients (MFCCs) and their first-order derivatives (MFCC_Ds) are calculated to represent the audio [Müller 2007]. Finally, CCA [Hardoon et al. 2004] is utilized to determine the STRs whose visual features have the best correlation with audio features. Slightly different from Izadinia et al. [2013], we use a state-of-the-art video segmentation approach [Xu et al. 2012], and we modify the final localization process to remove suspicious tiny motions.

3.1.1. Visual Analysis. The purpose of visual analysis is to segment the video sequence into K supervoxels SV_k ($k = 1, \ldots, K$ is the supervoxel index) and represent each supervoxel with some motion features. In frame F_t, pixels belonging to SV_k can be denoted as $SV_k(t)$ ($t = 1, \ldots, T$ is the frame index). We use a graph-based streaming hierarchical method [Xu et al. 2012] to segment the video sequence. Compared with the intraframe segmentation and interframe clustering method in Izadinia et al. [2013], the adopted video segmentation approach [Xu et al. 2012] is more robust and can get good results in various kinds of scenes. Motion features are velocity and acceleration derived from optical flows [Liu 2009]:

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

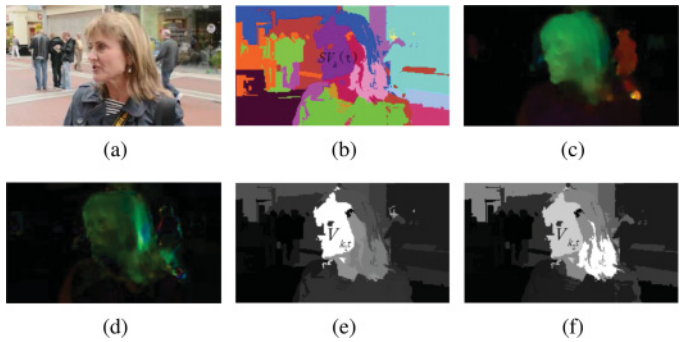

Fig. 3. Visual features for sound source localization. (a) Frame image $F_t, t = 119$ in this example. This video sequence shows an interviewee who is speaking. (b) Supervoxels in the current frame. $SV_k(t)$ denotes the face region supervoxel in F_t. (c) **vel**, velocity of F_t. (d) **acl**, acceleration of F_t. (e), (f) Mean velocity/acceleration magnitude of each supervoxel in F_t. $v_{k_1 t}$ and $v_{k_2 t}$ denote velocity and acceleration features of $SV_k(t)$.

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

where $\mathbf{U}^{+}(\mathbf{p}, t)$ represents forward optical flow from frame F_t to F_{t+1}; $\mathbf{U}^{-}(\mathbf{p}, t)$ represents backward optical flow from frame F_t to F_{t-1}; $\mathbf{p} = (i, j)$ denotes pixel location. Then, $SV_k(t)$ can be described by the mean velocity and acceleration magnitude of pixels belong to $SV_k(t)$. According to the variances along the time axis, we select the most dominant m_1 supervoxels for velocity and m_2 supervoxels for acceleration. Finally, we can use matrix $\mathbf{v}$ to characterize the whole video sequence:

$$\mathbf{v} = (v_{kt})_{M \times T} = [\mathbf{v}_1, \ldots, \mathbf{v}_T], \tag{3}$$

where v_{kt} denotes the visual feature (mean velocity or acceleration magnitude) of $SV_k(t)$; $\mathbf{v}_t, t = 1, \ldots, T$ is an $M = m_1 + m_2$ dimension vector that denotes the visual features of frame F_t. M supervoxels are selected according to the variance of $(v_{k\cdot})_{1 \times T}$. Figure 3 illustrates an example of velocity, acceleration, and corresponding derived visual features. SV_k indicates the supervoxel of the face region in this example. Then, $v_{k_1 t}$ and $v_{k_2 t}$ denote the velocity and acceleration of $SV_k(t)$. Features of other supervoxels are calculated similarly. Parameters settings will be given in Section 4.2.

3.1.2. Audio Analysis. We assume that the audio signal is dominated by the sound emitting from the target moving–sound-generating object. We extract $N/2$ MFCCs and $N/2$ first-order derivatives (MFCC_Ds) as audio features. Before extracting features, the audio signal is first framed to have the same number of frames as the video sequence. Then, the audio signal can be characterized by matrix $\mathbf{a} = [\mathbf{a}_1, \ldots, \mathbf{a}_T]$, where $\mathbf{a}_t, t = 1, \ldots, T$ is an N dimension vector used to feature the t^{th} windowed audio signal. More details about the parameters will be given in Section 4.2.

3.1.3. Audiovisual Correlation. The goal of audiovisual correlation is to detect the supervoxels, that is, 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. CCA is a classic, yet efficient, method that can perform correlation analysis after project signals of different modality to a common coordinate system. 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 $\mathbf{w}_a^T \mathbf{a}$ [Hardoon et al. 2004]:

$$(\mathbf{w}_v, \mathbf{w}_a) = \arg \max_{\mathbf{w}_v, \mathbf{w}_a} \mathtt{Corr}(\mathbf{w}_v^T \mathbf{v}, \mathbf{w}_a^T \mathbf{a}). \tag{4}$$

Fig. 4. Results of audiovisual correlation analysis. (a) Frame image. This video sequence presents a singing man. (b) Supervoxels in the current frame. (c) Correlation map. (d) Motion map. (e) Modified map. (f) Localization result, that is, audio attention map S_a.

Equation(4) has a closed-form solution as an eigenvalue problem:

$$\begin{cases} C_{vv}^{-1}C_{va}C_{aa}^{-1}C_{av}\mathbf{w}_v = \rho^2\mathbf{w}_v \\ C_{aa}^{-1}C_{av}C_{vv}^{-1}C_{va}\mathbf{w}_a = \rho^2\mathbf{w}_a \end{cases}, \tag{5}$$

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 Equation (4) is equivalent to finding the largest eigenvalue (denoted by ρ_1^2) and corresponding eigenvectors (denoted by $\mathbf{w}_{v^1}, \mathbf{w}_{a^1}$) in Equation (5). 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, that is, maximum eigenvalue ρ_1^2.

We generate a correlation map according to $\mathbf{w}_{v^1}$. Normalized components of $\mathbf{w}_{v^1}$ larger 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. In Izadinia et al. [2013], the correlation map is generated by assigning value 1 to pixels of all candidate supervoxels. Our method is more reasonable since better correlation denotes higher possibility of a position being the sound source. Moreover, our correlation map is further modified by multiplying a motion map to remove tiny motions. In the 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 spatiotemporally smoothed to the final localization result, which is taken as the audio attention map S_a.

Figure 4 illustrates some results of audiovisual correlation analysis. The analysis is in the supervoxel level. From Figure 4(c), we can see that some supervoxels with tiny motions sometimes are highly correlated with the audio. Thus, the proposed multiplication operation is essential. The effectiveness of the modification process is evident from Figure 4(c) and Figure 4(e).

3.2. Visual-Attention Modeling

Since the main purpose of this work is to demonstrate the superiority of audiovisual attention fusion, we model visual attention using state-of-the-art saliency algorithms. The performance improvement for different models will be compared. Both image and video saliency models are considered. Figure 5 demonstrates the framework of visual-attention modeling and audiovisual attention fusion. For the image-saliency model, we predict visual attention from both spatial and temporal aspects. Then the spatial, temporal, and audio attention maps are fused. For the video saliency model, the

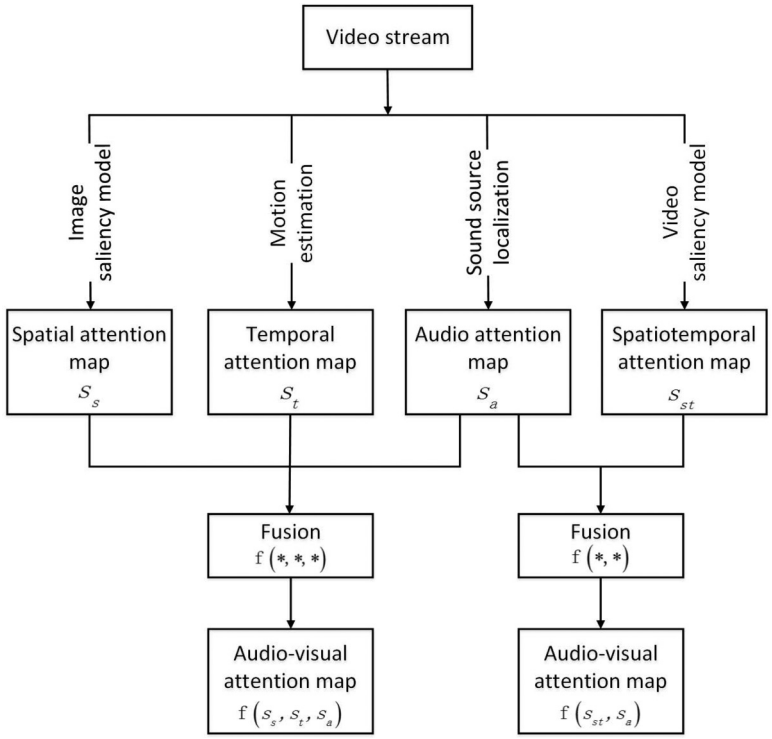

Fig. 5. Visual-attention modeling and audiovisual-attention fusion.

spatiotemporal attention maps are directly calculated from the video-saliency models. The spatiotemporal attention maps are then fused with the audio-attention maps.

3.2.1. Spatial-Attention Modeling. In recent years, dozens of saliency models have achieved good performances for static images. We choose 15 typical image-saliency algorithms to model spatial attention: IT [Itti et al. 1998], GBVS [Harel et al. 2006], SR [Hou and Zhang 2007], SUN [Zhang et al. 2008], PFT [Guo et al. 2008], SMVJ [Cerf et al. 2008], Hou [Hou and Zhang 2009], SeR [Seo and Milanfar 2009], Judd [Judd et al. 2009], FT [Achanta et al. 2009], RC [Cheng et al. 2011], BMS [Zhang and Sclaroff 2013], CovSal [Erdem and Erdem 2013], HFT [Li et al. 2013], and FES [Gu et al. 2015a]. Most are fixation prediction models, except for FT and RC. These 2 models were originally developed to detect salient objects. Note that several models take some middle or high-level features into consideration. For example, GBVS considers center bias; SMVJ includes a face detector; Judd considers more high-level factors, such as faces, persons, and cars.

The PFT method can be extended to PQFT [Guo et al. 2008], which considers motion features and works for video sequences. We consider only PFT in this section. PQFT will be considered in Section 3.2.3. Similarly, SeR method works for both image and video sequences [Seo and Milanfar 2009]; we consider only the image version here. The spatial-attention map is denoted as S_s in following sections.

3.2.2. Temporal-Attention Modeling. Object motion is an important cue for visual attention. Optical flow is often used to describe the local motion of video sequences [Liu 2009]. For temporal attention, we adopt optical flow–based motion estimation to reduce computation since we have calculated optical flow for each frame in the process

of audio-attention modeling. The temporal attention map S_t can be calculated by

$$S_t = g * ||\mathbf{vel}||, \tag{6}$$

where g is a Gaussian kernel, $*$ is the convolution product operator, and **vel** is the velocity acquired by Equation (1).

3.2.3. Spatiotemporal-Attention Modeling. Although most image saliency models work for video sequences, some researchers make special considerations for video sequences. Guo et al. [2008] extended their PFT method to PQFT by considering both spatial and motion features. Seo and Milanfar [2009] proposed a unified framework for both static and space–time saliency detection. The proposed method calculated saliency using a "self-resemblance" measure, and it did not require explicit motion estimation by taking the video sequence as a volume. Kim et al. [2015] recently presented a method incorporating spatial and temporal features into a random walk with restart (RWR) framework to detect spatiotemporal saliency (RWRV). All these methods will be considered and used to detect spatiotemporal-attention maps. The saliency map computed by video saliency algorithm is denoted as S_{st} in following contents.

3.3. Audiovisual-Attention Fusion

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

$$\begin{aligned} S &= f(S_s, S_t, S_a) \text{ or} \\ S &= f(S_{st}, S_a), \end{aligned} \tag{7}$$

where S is the final audiovisual saliency map; S_s, S_t, S_{st}, S_a are spatial-, temporal-, spatiotemporal-, and audio-attention maps, respectively; and f is the fusion function. Chamaret et al. [2010] evaluated common spatial and spatiotemporal fusion strategies. In this work, we test the following 3 classical and most commonly used fusion methods:

—Normalization and summation (NS):

$$f : S_i \to \mathcal{N}\Big(\sum_i \mathcal{N}(S_i)\Big) \tag{8}$$

—Normalization and max (NM):

$$f : S_i \to \mathcal{N}\Big(\max_i \mathcal{N}(S_i)\Big) \tag{9}$$

—Normalization and product (NP):

$$f : S_i \to \mathcal{N}\Big(\prod_i \mathcal{N}(S_i)\Big), \tag{10}$$

where, in Equations (8), (9), and (10), $i \in \{s, t, a\}$ or $\{st, a\}$; and $\mathcal{N}$ is a normalization operator to normalize all attention maps to the same dynamic range, that is, $[0, 1]$. We mainly use NS as our fusion method because it is more intuitive. Moreover, the NS method can achieve state-of-the-art performance, which is discussed in Section 4.4. The performance of the other two methods are also compared in that section.

4. EXPERIMENTS AND RESULTS

4.1. Eye-tracking Experiments

4.1.1. Stimuli. Few audiovisual-attention databases are made available in the community now. Coutrot and Guyader [2013] constructed an audiovisual-attention database,

Table I. Main Specifications and Contents of Test Video Sequences

Index	Video sequence	Resolution W×H	Frame rate fps	No. of frames	Video content	Audio content
V1	Speech1	1280×720	29.97	255	Three people's talk	One person's speech
V2	Speech2	1280×720	29.97	240	Speaking	Speaking
V3	Speech3	1280×720	29.97	243	Speaking	Speaking
V4	Interview1	800×450	25	225	Interviewer and pedestrians	Interviewer's speech
V5	Interview2	1280×720	24	152	Interviewee and pedestrians	Interviewee's speech
V6	Basketball1	1280×720	23.98	193	Basketball dribble	Basketball dribble
V7	Basketball2	1280×720	23.97	240	Basketball dribble	Basketball dribble
V8	News1	1280×720	25	255	Two people's news report	One reporter's speech
V9	News2	1280×720	25	125	Two people's news report	One reporter's speech
V10	News3	640×356	29.97	308	News report	Reporter's speech
V11	News4	1280×720	25	125	Two people's news report	One reporter's speech
V12	Conservation1	1024×576	25	145	Conservation	One person's speech
V13	Conservation2	1280×720	25	175	Conservation	One person's speech
V14	Conservation3	1280×720	29.97	265	Conservation	One person's speech
V15	Drummer1	1280×720	25	150	Drummer and crowd	Drumbeat
V16	Drummer2	960×540	30	240	Drummer and pedestrians	Drumbeat
V17	Soccer1	1280×720	30	180	Pop and tip	Sound of pop and tip
V18	Soccer2	1280×720	29.97	150	Pop and tip	Sound of pop and tip
V19	Singing1	1280×720	29.97	240	Singing	Singing
V20	Singing2	1280×720	29.97	271	Singing	Singing
V21	Tap1	960×540	29.97	241	Two tap dancers	One dancer's tap dancing
V22	Tap2	800×450	30	300	Dancing	Dancing
V23	Tap3	1280×720	30	240	Tap dancer in the crowd	The tap dancing sound
V24	Tap4	1280×720	25	200	Two tap dancers in the crowd	The tap dancing sound
V25	Piano1	1280×720	30	300	Piano player and crowd	Piano
V26	Piano2	1280×720	29.97	150	Piano player and crowd	Piano
V27	Piano3	1280×720	29.97	301	Piano player and crowd	Piano
V28	Dog1	1280×720	29.97	210	Dog's running and barking	Dog's bark
V29	Dog2	1280×720	30	150	A dog is barking at a toy	Dog's bark
V30	Dog3	1280×720	30	210	A dog is barking at a toy	Dog's bark
V31	Bird	1280×720	29.97	150	Bird	Twitter of the bird
V32	Dancers	1280×720	28.67	202	Two dancers	The music rhythm
V33	Harp	1280×720	29.97	301	Harp player	Sound of the harp
V34	Beat	1280×720	30	300	Instrumental beat	Instrumental beat
V35	Squirrel	1280×720	30	210	Squirrel	Sound of the squirrel
V36	Guitar1	640×360	23.98	156	Playing guitar	Guitar sound
V37	Guitar2	640×360	25	126	Guitar player and pedestrians	Guitar sound
V38	Guitar3	480×360	29.97	251	Guitar player and attendant	Guitar sound
V39	Guitar4	1280×720	29.97	240	Guitar player and pedestrians	Guitar sound
V40	Guitar5	1280×720	25	200	Guitar player and pedestrians	Guitar sound
V41	Violin1	320×240	25	129	Playing violin	Violin sound
V42	Violin2	1280×720	25	250	Playing violin in the crowd	Violin sound
V43	Violin3	1280×960	23.98	190	Playing violin	Violin sound
V44	Darbuka1	1280×720	29.97	180	Darbuka player and crowd	Darbuka playing
V45	Darbuka2	720×720	25	200	Darbuka player and pedestrian	Darbuka playing

but the database was originally built to investigate the mechanism affecting how sound impacts eye movements. The presented scenes are too general, and are not applicable in this work since our methods are mainly used in scenes in which the motion and audio are highly correlated. In this article, we perform eye-tracking experiments with 45 test video sequences. Several video sequences have been used in Izadinia et al. [2013] and Li et al. [2014]; the rest are gathered from YouTube. Both video sequences and eye-tracking data will be publicly available. Main specifications of the collected video sequences are listed in Table I. We also list the brief descriptions of both video

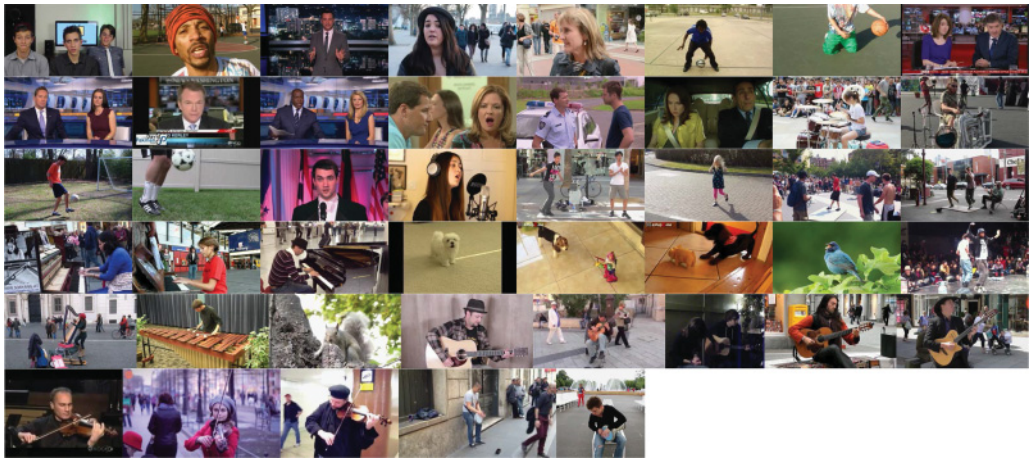

Fig. 6. Thumbnails of all test video sequences. A total of 45 video sequences were used. From left to right, from top to bottom: V1,V2,. . .,V45.

and audio contents in this table for better understanding of the video sequences. The lengths of video sequences range from 5s to 10s. A total of 9564 frames are used in our experiments. Sample frames of all test video sequences are illustrated in Figure 6. From this figure, we can see that collected videos contain a variety of scenes, for example, dancing, playing musical instruments, playing or kicking the ball, speaking, and talking. All video sequences contain soundtracks, and most video sequences only contain one dominant sound that is emitted by the main process of video sequences. Some video sequences contain disturbed motion that emits no sound, such as pedestrians' passing by and the motion of conversational partners.

4.1.2. Apparatus. We use Tobii T120 Eye Tracker to collect eye movement data. Tobii T120 has a 17in screen, whose resolution is 1280 $\times$ 1024 pixels. It has an effective tracking range of 50$\sim$80cm. Subjects are seated around 60cm from the eye tracker. The sampling rate is set to 120Hz.

4.1.3. Participants. A total of 16 inexperienced college students participated in our experiments. Subjects were not familiar with the purpose of the experiments, and they all had normal or corrected-to-normal vision. All subjects watched all 45 test videos. Two subjects' eye-movement data were abandoned because insufficient eye-movement data was tracked.

4.1.4. Procedure and Test Conditions. We adopted a free-viewing approach in the tests. Moreover, eye-tracking experiments were conducted with an audiovisual test condition. Subjects were told to just watch the video sequences and listen to the soundtrack played through the headset. We performed a 5-point calibration at the beginning of the tests. After that, subjects looked at the center of the screen and the tests began. During the tests, video sequences were linearly rescaled to fit the maximum resolution of the screen, but we did not change the aspect ratio of video sequences. For example, a video sequence with the resolution of 320$\times$240 was spatially interpolated to the resolution of 1280$\times$960. We adopted the bicubic interpolation method in the experiments. All video sequences were played in a random order, and the presenting order for each subject was also different. Between two video sequences, there was a gray-screen interval of 2s.

Table II. 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

4.2. Implementation Details

In the implementation, all video sequences were analyzed at their original frame rate. To reduce computation, we downsampled the spatial resolution of video sequences to make them have a maximum width or height of 240 pixels without change of the video sequences' aspect ratio. In the final saliency evaluation stage, however, saliency maps were linearly enlarged to the resolution in which the eye-movement data was tracked. Before audio processing, the audio signal was framed to have the same number of frames as the video sequence, and the framing windows were 50% overlapped Hamming windows. The adopted video segmentation approach [Xu et al. 2012] is a hierarchical method. We chose the desired level so that the final number of supervoxels was most close to 25. More model implementation parameters can be found in Table II. Note that parameters are mainly from third-party algorithms, like video segmentation [Xu et al. 2012] and optical flow estimating [Liu 2009], and we rarely tuned these parameters.

4.3. Evaluation Metrics

Similar to the study by Borji et al. [2013], we used 3 saliency evaluation metrics, but we replaced the original area under the Receiver Operating Characteristic (ROC) curve (AUC) with the shuffled version of AUC (sAUC) to wipe off the influence of center bias [Zhang et al. 2008].

—*sAUC*: The saliency map acts as a binary classifier. Values greater than a threshold were classified as fixated, while the rest were classified as nonfixated. Human eye fixations were ground-truth fixated positions, whereas the same number of random locations sampled uniformly from fixations of all other images were taken as non-fixated positions. Then, the true-positive rate and the false-positive rate could be calculated. As the threshold varied, the ROC curve was plotted. The area under the ROC curve indicates how well saliency map predicts eye fixations.
—*Linear Correlation Coefficient (CC)*: It simply calculates the 2D correlation coefficient between the saliency map and fixation density map (FDM), for which the FDM is generated by convolving the human eye-fixation map with a Gaussian kernel.

Table III. Performance of Saliency Algorithms and Their Combination with Temporal and Audio-Attention Maps

	sAUC				CC				NSS			
Model	S	ST	STA	Sig	S	ST	STA	Sig	S	ST	STA	Sig
IT	0.6178	0.7003	0.7431	+1	0.3166	0.3758	0.4264	+1	1.264	1.556	1.805	+1
GBVS	0.6529	0.7165	0.7485	+1	**0.3524**	**0.3971**	**0.4422**	+1	**1.454**	**1.666**	**1.886**	+1
SR	0.6635	**0.7244**	**0.7625**	+1	0.2409	0.3389	0.4107	+1	1.008	1.436	1.765	+1
SUN	0.6037	0.7051	0.7526	+1	0.1661	0.3029	0.3878	+1	0.695	1.292	1.674	+1
PFT	0.6423	0.7141	0.7554	+1	0.2015	0.3265	0.4045	+1	0.850	1.385	1.740	+1
SMVJ	**0.6665**	0.7211	0.7505	+1	**0.3690**	**0.4068**	**0.4472**	+1	**1.525**	**1.706**	**1.905**	+1
Hou	0.5862	0.7072	0.7479	+1	0.1848	0.3343	0.4077	+1	0.845	1.458	1.784	+1
SeR	0.6281	0.7003	0.7475	+1	0.1982	0.2971	0.3766	+1	0.818	1.247	1.607	+1
Judd	0.6594	0.7236	0.7529	+1	0.3192	0.3815	0.4305	+1	1.351	1.630	1.859	+1
FT	0.5087	0.6666	0.7301	+1	0.0386	0.2609	0.3654	+1	0.156	1.109	1.583	+1
RC	0.5862	0.6836	0.7378	+1	0.2134	0.3260	0.3985	+1	0.847	1.363	1.704	+1
BMS	**0.6875**	**0.7373**	**0.7626**	+1	0.2684	0.3608	0.4203	+1	1.273	1.608	1.856	+1
CovSal	0.6207	0.7048	0.7426	+1	**0.4019**	**0.4295**	**0.4612**	+1	**1.686**	**1.812**	**1.975**	+1
HFT	0.6383	0.7080	0.7493	+1	0.3365	0.3852	0.4355	+1	1.404	1.625	1.865	+1
FES	**0.6842**	**0.7372**	**0.7672**	+1	0.2341	0.3331	0.4031	+1	1.003	1.427	1.742	+1
Mean	0.6297	0.7100	0.7500	+1	0.2561	0.3504	0.4145	+1	1.079	1.488	1.783	+1
PQFT	-	0.6663	0.7416	+1	-	0.2603	0.3995	+1	-	1.116	1.746	+1
SeR	-	0.5920	0.7186	+1	-	0.2267	0.3902	+1	-	0.968	1.706	+1
RWRV	-	0.6246	0.7288	+1	-	0.2403	0.3833	+1	-	0.981	1.634	+1
Mean	-	0.6276	0.7297	+1	-	0.2424	0.3910	+1	-	1.022	1.695	+1

*S: Spatial; ST: Spatiotemporal; STA: Spatiotemporal-Audio. We highlight three top-performed models in each column. Sig: statistical significance comparison between STA and ST; "+1" denotes STA is statistically better than ST. Note that all models perform significantly better after incorporating audio attention.

—*Normalized Scanpath Saliency (NSS)*: The mean value of the normalized saliency map at all fixation points is calculated as NSS [Peters et al. 2005], for which the normalized saliency map have zero mean and unit standard deviation.

For all 3 metrics, greater values denote better consistency between the predicted saliency map and the ground-truth eye-tracking data. We used the evaluation code downloaded from Bylinskii et al. [2012].

4.4. Results and Analysis I: Effectiveness of Incorporating Audio Cues

In this section, we demonstrate the effectiveness of incorporating audio cues based on the gathered video sequences and eye-movement data. For spatial saliency models, three types of attention maps were evaluated: S_s, $f(S_s, S_t)$, and $f(S_s, S_t, S_a)$. For spatiotemporal models, two kinds were evaluated: S_{st} and $f(S_{st}, S_a)$. As already defined, S_s, S_t, S_{st}, S_a are spatial-, temporal-, spatiotemporal-, and audio-attention maps, respectively; f is the fusion function. In this article, saliency maps calculated from 15 image algorithms and 3 video algorithms act as S_s and S_{st}, respectively.

Experiment results are listed in Table III, in which S, ST, and STA denote the performance of spatial (S_s)-, spatiotemporal ($f(S_s, S_t)$ or S_{st})- and audiovisual ($f(S_s, S_t, S_a)$ or $f(S_{st}, S_a)$)-attention maps, respectively. In this table, listed results are average performance of all video frames. Note that we only list the results fused by the NS method. Other fusion methods are compared later. If only considering spatial information, some recently proposed models (such as BMS [Zhang and Sclaroff 2013], CovSal [Erdem and Erdem 2013], and FES [Gu et al. 2015a]), and several classical models incorporating high-level factors (such as GBVS [Harel et al. 2006], SMVJ [Cerf et al. 2008], and Judd [Judd et al. 2009]) generally perform better. FT [Achanta et al. 2009] and RC [Cheng

Table IV. Comparison of Different Fusion Methods

Fusion method f		NS	NM	NP
sAUC	S_s	0.6297		
	$f(S_s, S_t)$	0.7100	0.6886	0.6820
	$f(S_s, S_t, S_a)$	0.7500	0.7172	0.6747
CC	S_s	0.2561		
	$f(S_s, S_t)$	0.3504	0.3096	0.3535
	$f(S_s, S_t, S_a)$	0.4145	0.3390	0.3924
NSS	S_s	1.0786		
	$f(S_s, S_t)$	1.4878	1.3123	1.5839
	$f(S_s, S_t, S_a)$	1.7833	1.4355	1.9141

et al. 2011] have lower performances. It is expected since they are originally designed to detect salient objects.

For most tested saliency models, not surprisingly, $f(S_s, S_t)$ performs better than S_s since motion is an important incentive for visual attention. $f(S_s, S_t, S_a)$ (or $f(S_{st}, S_a)$) performs even better than $f(S_s, S_t)$ (or S_{st}), however. It is a quantitative verification of our framework. Using $p(\cdot)$ to denote the performance of the saliency map, the following inequality holds:

$$\begin{aligned} &p(f(S_s, S_t, S_a)) > p(f(S_s, S_t)) > p(S_s) \text{ or} \\ &\quad p(f(S_{st}, S_a)) > p(f(S_{st})). \end{aligned} \tag{11}$$

The performance enhancement exists no matter what evaluation metrics are used, that is, $p(\cdot)$ can be sAUC, CC, or NSS scores. Significance tests, one-way analysis on variance (ANOVA) [Snecdecor and Cochran 1989], are performed to verify if performance promotion is significant. Results show that all models improved significantly ($p < 0.001$) after combining with audio-attention maps. Thus, we can reach the conclusion that audio has some influence on visual attention and we can promote human fixation prediction by incorporating audio cues.

Figure 7 is an intuitive illustration of related saliency maps. For each sample video sequence, frame image, fixation density map and different kinds of attention maps mentioned earlier are shown. Six video sequences are selected as examples in this figure. In the selected video sequences, S_a shows good correlation with the FDM, and the final saliency map becomes better after incorporating S_a. Specifically, taking V13 as an example, it is a typical case 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 $f(S_s, S_t)$. Other video sequences behave similarly, which can be seen from Figure 7. Purely spatial and temporal analysis may not predict visual attention perfectly. Through audiovisual analysis, howeve, we can locate and emphasize the sound-emitting regions that draw great visual attention, thus providing better results.

Different fusion methods are also compared. Table IV lists the results, which are the average performance of all video frames and all original saliency algorithms. Fusion methods described in Section 3.3 are tested and listed in this table. Note that S_s does not require any fusion process; we include it here for the convenience of performance comparison. According to Table IV, no matter under what kind of evaluation metrics and using which fusion methods, inequality Equation (11) generally holds. Since the NS method can achieve state-of-the-art performance, we mainly analyze the results derived by the NS fusion method in this article.

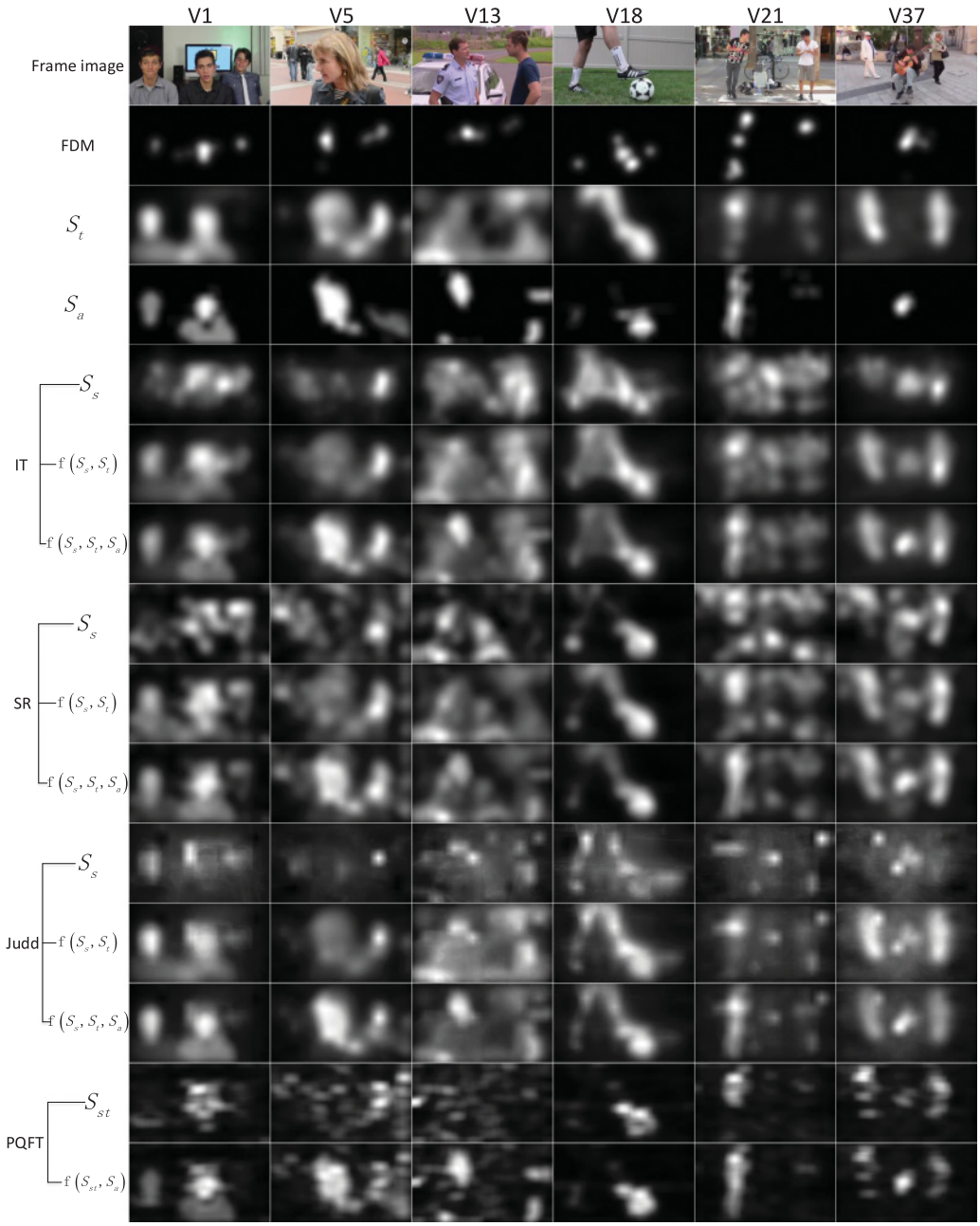

Fig. 7. Examples of related saliency maps. One column corresponds to one video sequence, and the first column lists the content names of corresponding rows. Six video sequences are chosen as examples in this figure. For each video sequence, example frame, FDM, motion map, audio-attention map and saliency maps computed and fused from several original algorithms are illustrated.

Table V. Performance Gain of Test Video Sequences

Video	G_t	G_{ta}	G_a	Video	G_t	G_{ta}	G_a	Video	G_t	G_{ta}	G_a
V1	48.45%	75.96%	17.47%	V16	31.41%	41.06%	6.67%	V31	9.46%	18.13%	7.66%
V2	47.71%	70.86%	14.78%	V17	122.02%	106.98%	−5.85%	V32	60.75%	78.26%	9.72%
V3	9.84%	−2.92%	−11.30%	V18	31.13%	54.44%	16.77%	V33	6.08%	0.24%	−5.51%
V4	36.99%	94.67%	40.26%	V19	73.69%	97.36%	12.33%	V34	46.86%	47.18%	−0.05%
V5	82.13%	199.90%	57.33%	V20	27.14%	40.39%	10.07%	V35	−3.75%	46.65%	52.06%
V6	28.68%	41.99%	10.02%	V21	75.45%	146.67%	38.26%	V36	13.67%	22.10%	7.40%
V7	−7.27%	−1.52%	6.28%	V22	16.60%	23.55%	5.80%	V37	−9.33%	11.73%	24.37%
V8	93.76%	250.13%	72.54%	V23	18.30%	31.64%	11.13%	V38	47.16%	77.27%	18.70%
V9	49.70%	91.08%	25.47%	V24	73.52%	80.76%	3.23%	V39	75.62%	140.40%	32.91%
V10	123.66%	172.86%	19.07%	V25	109.61%	224.64%	49.09%	V40	1.55%	21.47%	19.62%
V11	46.82%	−10.24%	−37.18%	V26	16.67%	52.48%	30.04%	V41	−9.16%	−6.02%	3.52%
V12	80.84%	125.02%	23.60%	V27	11.95%	25.32%	11.90%	V42	5.67%	35.18%	27.82%
V13	−8.36%	49.86%	65.79%	V28	−9.31%	7.90%	20.09%	V43	−3.05%	4.97%	8.35%
V14	108.36%	224.24%	49.56%	V29	26.89%	44.78%	13.33%	V44	17.18%	−11.75%	−23.78%
V15	61.78%	53.14%	−4.62%	V30	38.62%	57.51%	12.80%	V45	46.54%	78.70%	20.26%

4.5. Results and Analysis II: Mechanism of the Promotion Effect of Multimodal Analysis

In this section, we analyze the performance gain caused by the added audio information for all test video sequences. We provide some hints regarding when and how audiovisual analysis benefits visual-attention prediction. In detail, using G to denote the performance gain, we analyze the following results of all test video sequences:

—Performance gained from motion information:

$$G_t = \frac{p(f(S_s, S_t)) - p(S_s)}{p(S_s)} \tag{12}$$

—Performance gained from motion and audio information:

$$G_{ta} = \frac{p(f(S_s, S_t, S_a)) - p(S_s)}{p(S_s)} \tag{13}$$

—Performance gained from audio information:

$$G_a = \frac{p(f(S_s, S_t, S_a)) - p(f(S_s, S_t))}{p(f(S_s, S_t))}, \tag{14}$$

where $p(\cdot)$ indicates the performance of saliency map.

Table V lists the results. It should be noted that the performance gain G in this table is averaged over all 3 evaluation metrics. Moreover, we consider all 15 image saliency algorithms, and what is listed in Table V are average results. In addition to the quantitative results, we provide some illustrations of example attention maps in Figure 8, which is a supplement to Figure 7. These two figures together provide at least one example frame image, corresponding FDM, spatial-, temporal-, audio-, and final audiovisual-attention maps for each test video sequence.

The effectiveness of motion information is evident from G_t. Performance of most video sequences improves substantially except for several video sequences. As shown in Figure 8, primary motion may not represent the focus of visual attention exactly in these video sequences (V7, V13, V37, V41). It may also be caused by camera motion (V28, V35, V43). As is for audio, if $G_{ta} > G_t$ or $G_a > 0$, audio information helps. Through analysis, we find that audiovisual analysis contributes much to fixation prediction generally in the following scenarios:

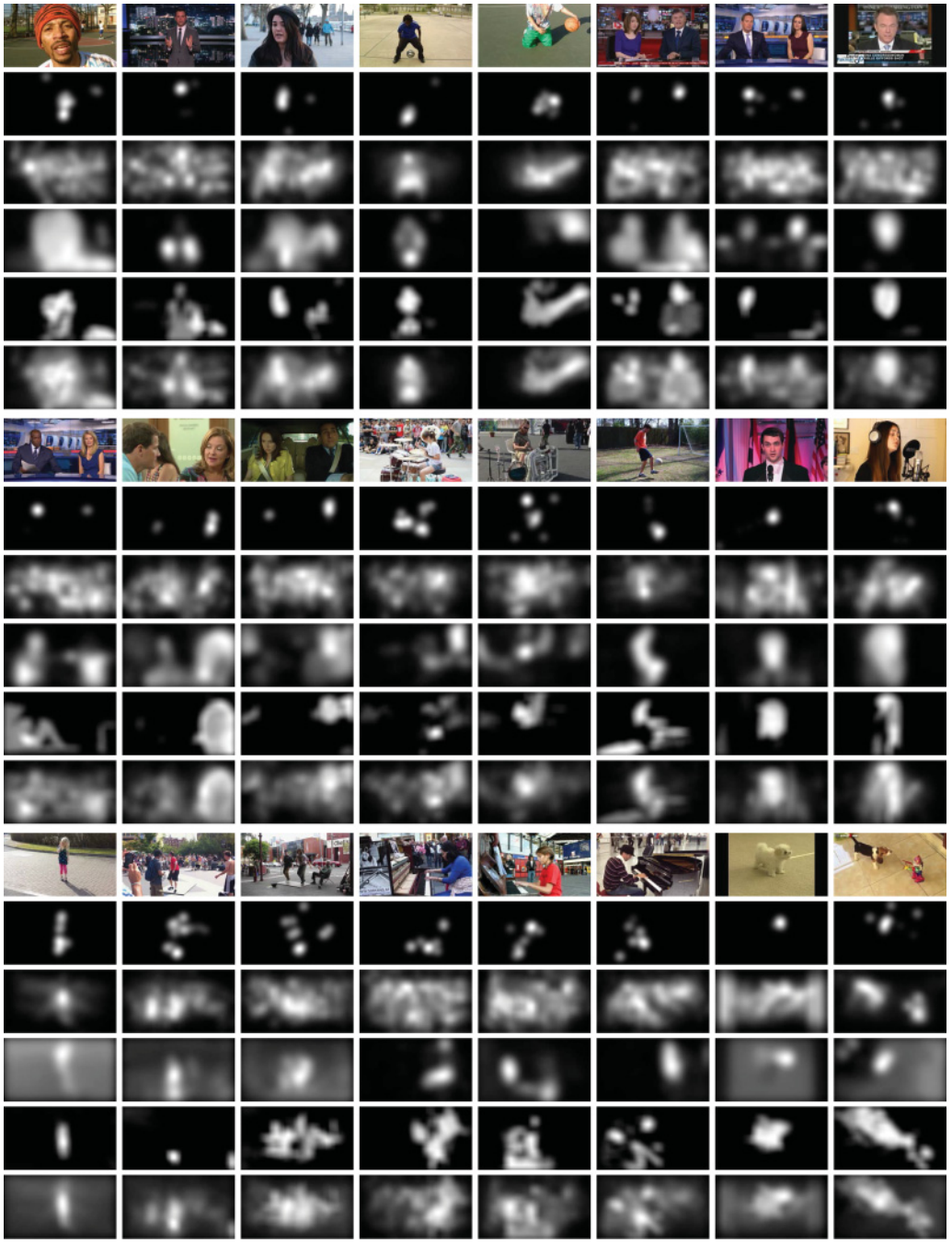

Fig. 8. Saliency maps of videos except what has been shown in Figure 7. For each video, 6 images are shown (from top to bottom): frame image, FDM, S_s, S_t, S_a, $f(S_s, S_t, S_a)$. S_s are computed by the IT method [Itti et al. 1998] in this figure. The first row (frame image, from left to right): V2, V3, V4, V6, V7, V8, V9, V10. The seventh row: V11, V12, V14, V15, V16, V17, V19, V20. The thirteen row: V22, V23, V24, V25, V26, V27, V28, V29.

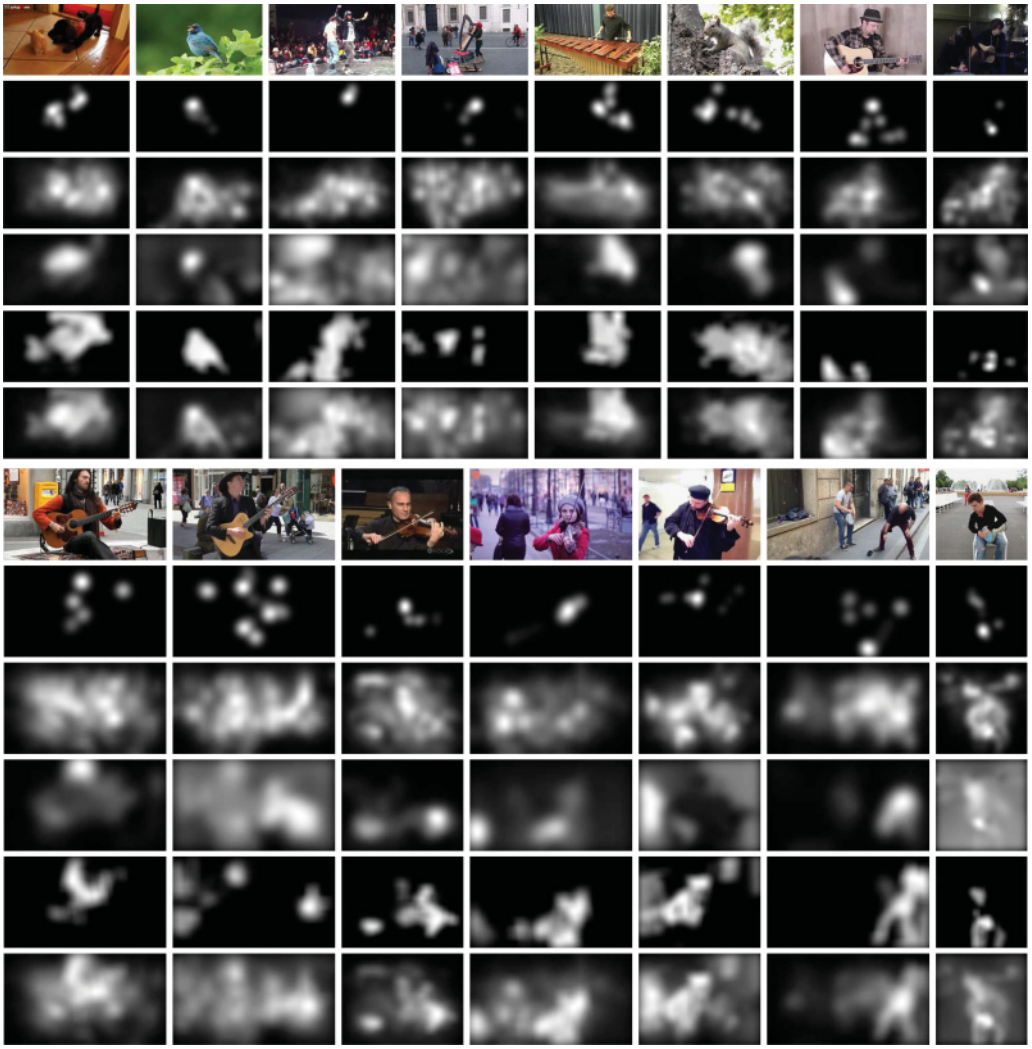

Fig. 8. (continued) Saliency maps of videos except what has been shown in Figure 7. The first row: V30, V31, V32, V33, V34, V35, V36, V38. The thirteenth row: V39, V40, V41, V42, V43, V44, V45.

—There are several salient objects or targets in the scene, but only one or part of them are emitting sound. In such circumstances, subjects may notice all salient targets. Since one target is generating sound, however, this target is relatively more salient compared with others. Thus, subjects pay more attention to the sound source. Through multimodal analysis, we can locate and emphasize the sound source, thus providing better performance. V1, V8, V9, V12, V13, V14, V21, V23, V37, and V38 are typical scenes of this kind. For example, as already analyzed in Section 4.4, S_a helps a lot in V13.

—In the video sequence, only one salient target exists, and is moving and generating sound. There is some disturbed motion, however, especially in the background regions, or the video sequences may contain camera motion. In this kind of scene, the moving–sound-generating object is both visually and aurally salient. Subjects will focus more on the moving–sound-generating object and rarely pay attention to such

kinds of disturbed motion. The traditional video saliency model suffers from these motion, however. Through sound source localization, we can reduce the influence of disturbed motion. V4, V5, V16, V22, V24, V25, V26, V27, V28, V29, V30, V35, V39, V40, V42, V43, and V45 are representative video sequences. Taking V4 as an instance, S_a eliminates a lot of irrelevant motion, as shown in Figure 8. Thus, the final fused map works better.

—Similar to the second case, one main process exists, but there is no background motion. Audiovisual analysis will help to concentrate more on the sound-emitting region, which is always the focus of attention. V2, V6, V7, V10, V18, V19, V20, V31, V32, V36, and V41 all belong to this kind. For example, in V10, V19, and V20, we can locate the speaking or singing faces accurately. In these video sequences, the face regions are also areas with large movement. In other words, the sound-emitting regions and motion regions highly overlap. The enhancement effect of the audio-attention map is also very useful, however.

Admittedly, audiovisual analysis may not always be helpful in fixation prediction. Sometimes, the sound-emitting regions are not as attractive compared with other positions, in which case, highlighting the sound sources is useless. Also, the sound-source localization may provide results that are not as good. For instance, in V3, V11, and V44, $f(S_s, S_t, S_a)$ performs worse than $f(S_s, S_t)$. Thus, an accurate and robust sound source localization method is also indispensable. However, it is worth mentioning that the audio-attention map-generating process can be interpreted as a procedure that picks out and highlights a part of or the whole motion region. We do not need to worry that the audio-attention map will do much harm to the final saliency map since the localized regions generally have large movement. An interesting phenomenon is that the G_a of video sequences belonging to the first and second case are generally greater than that of the third case. As shown in Table V, G_a of V4, V5, V8, V13, V14, V21, V25, V26, V35, V39, and V42 are greater than the average level. This is reasonable because the motion picking and highlighting effect in video sequences of the the first and second case is more significant.

5. 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 audio information to human eye fixation prediction. Through audiovisual correlation analysis, we locate the moving–sound-generating objects, then generate an audio-attention map for each frame. The audio-attention maps are further fused with conventional visual-attention maps. The efficiency of generated audiovisual-saliency maps is verified with gathered video sequences and eye-movement data. Moreover, experiment results show that audio-attention map generating is a process of selecting and emphasizing regions whose motion is highly correlated with the audio. Thus, our approach is extremely useful in scenes containing moving–sound-generating objects and other disturbed motion that has no relation with the audio. Of course, incorporating audio information may not always contribute to eye-fixation prediction because of factors such as the accuracy of moving–sound-generating object localization and the attractiveness of sound sources. Thus, more robust and accurate sound-localization methods and comprehensive investigation of audio's role in visual-attention prediction are directions of future efforts.

REFERENCES

Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 1597–1604.

Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. 2012. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11, 2189–2202.

Xuan Bao and Romit Roy Choudhury. 2010. Movi: Mobile phone based video highlights via collaborative sensing. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services.* ACM, 357–370.

Zohar Barzelay and Yoav Y. Schechner. 2007. Harmony in motion. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 1–8.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2014. Salient object detection: A survey. *ArXiv Preprint.*

Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1, 185–207.

Ali Borji, Dicky N. Sihite, and Laurent Itti. 2013. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* 22, 1, 55–69.

Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2012. MIT Saliency Benchmark. Retrieved September 25, 2016 from http://saliency.mit.edu/.

Moran Cerf, E. Paxon Frady, and Christof Koch. 2009. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision* 9, 12, 10.

Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems.* 241–248.

Christel Chamaret, Jean-Claude Chevet, and Olivier Le Meur. 2010. Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. In *Proceedings of IEEE International Conference on Image Processing.* 1077–1080.

Yanxiang Chen, Tam V. Nguyen, Mohan Kankanhalli, Jun Yuan, Shuicheng Yan, and Meng Wang. 2014. Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 11, 1992–2003.

Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. 2011. Global contrast based salient region detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 409–416.

Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. 2014. BING: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 3286–3293.

Antoine Coutrot and Nathalie Guyader. 2013. Toward the introduction of auditory information in dynamic visual attention models. In *Proceedings of IEEE International Workshop on Image Analysis for Multimedia Interactive Services.* 1–4.

Antoine Coutrot and Nathalie Guyader. 2014. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision* 14, 8, 5.

Erkut Erdem and Aykut Erdem. 2013. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision* 13, 4, 11.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* 15, 7, 1553–1568.

Ke Gu, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang. 2015a. Visual saliency detection with free energy theory. *IEEE Signal Processing Letters* 22, 10, 1552–1555.

Ke Gu, Guofu Zhai, Xu Yang, Wensheng Zhang, and Chang Wen Chen. 2015b. Automatic contrast enhancement technology with saliency preservation. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 9, 1480–1494.

Chenlei Guo, Qi Ma, and Liming Zhang. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 1–8.

David Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16, 12, 2639–2664.

Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. In *Advances in Neural Information Processing Systems.* 545–552.

Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 1–8.

Xiaodi Hou and Liqing Zhang. 2009. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems.* 681–688.

Laurent Itti. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing* 13, 10, 1304–1318.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11, 1254–1259.

Hamid Izadinia, Imran Saleemi, and Mubarak Shah. 2013. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia* 15, 2, 378–390.

Lloyd A. Jeffress. 1948. A place theory of sound localization. *Journal of Comparative and Physiological Psychology* 41, 1, 35.

Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of IEEE International Conference on Computer Vision.* 2106–2113.

Einat Kidron, Yoav Y. Schechner, and Michael Elad. 2005. Pixels that sound. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 1. 88–95.

Einat Kidron, Yoav Y. Schechner, and Michael Elad. 2007. Cross-modal localization via sparsity. *IEEE Transactions on Signal Processing* 55, 4, 1390–1404.

Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. 2015. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing.* 24, 8 (Aug. 2015), 2552–2564.

Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi. 2011. Subjective quality evaluation of foveated video coding using audio-visual focus of attention. *IEEE Journal on Selected Topics in Signal Processing* 5, 7, 1322–1331.

Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. 2013. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 4, 996–1010.

Kai Li, Jun Ye, and Kien A. Hua. 2014. What's making that sound? In *Proceedings of ACM International Conference on Multimedia*. 147–156.

Ce Liu. 2009. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. Ph.D. Dissertation. Citeseer.

Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. 2005. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia* 7, 5, 907–919.

Xiongkuo Min, Guangtao Zhai, Zhongpai Gao, Chunjia Hu, and Xiaokang Yang. 2014. Sound influences visual attention discriminately in videos. In *Proceedings of IEEE International Workshop on Quality of Multimedia Experience.* 153–158.

Vicente P. Minotto, Claudio R. Jung, and Bowon Lee. 2014. Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs. *IEEE Transactions on Multimedia* 16, 4, 1032–1044.

Meinard Müller. 2007. *Information Retrieval for Music and Motion*. Vol. 2. Springer.

Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and D. Barbba. 2007. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Proceedings of IEEE International Conference on Image Processing*, Vol. 2. II–169–II–172.

David R. Perrott, Kourosh Saberi, Kathleen Brown, and Thomas Z. Strybel. 1990. Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics* 48, 3, 214–226.

Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research* 45, 18, 2397–2416.

Hae Jong Seo and Peyman Milanfar. 2009. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision* 9, 12, 15.

G. W. Snecdecor and W. G. Cochran. 1989. *Statistical Methods* (8th ed.). Iowa State University Press, Iowa City, IA.

Guanghan Song, Denis Pellerin, and Lionel Granjon. 2013. Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research* 6, 4, 1–13.

Jean Vroomen and Beatrice de Gelder. 2000. Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance* 26, 5, 1583.

Chenliang Xu, Caiming Xiong, and Jason J. Corso. 2012. Streaming hierarchical video segmentation. In *European Conference on Computer Vision*, Springer, 626–639.

Steven Yantis and John Jonides. 1990. Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance* 16, 1, 121.

Yun Zhai and Mubarak Shah. 2006. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of ACM International Conference on Multimedia*. 815–824.

Jianming Zhang and Stan Sclaroff. 2013. Saliency detection: A Boolean map approach. In *Proceedings of IEEE International Conference on Computer Vision*. 153–160.

Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. 2008. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision* 8, 7, 32.

Received November 2015; revised July 2016; accepted August 2016