Learning to Predict Salient Faces: A Novel Visual-Audio Saliency Model

Yufan Liu[1,2†*] Minglang Qiao[4†] Mai Xu[4‡] Bing Li[1‡] Weiming Hu[1,2,3] Ali Borji[5]

[1]National Laboratory of Pattern Recognition, CASIA
[2]University of Chinese Academy of Sciences [3]CEBSIT
[4]Beihang University & Hangzhou Innovation Institute, Beihang University
[5]MarkableAI Inc.

Abstract. Recently, video streams have occupied a large proportion of Internet traffic, most of which contain human faces. Hence, it is necessary to predict saliency on multiple-face videos, which can provide attention cues for many content based applications. However, most of multiple-face saliency prediction works only consider visual information and ignore audio, which is not consistent with the naturalistic scenarios. Several behavioral studies have established that sound influences human attention, especially during the speech turn-taking in multiple-face videos. In this paper, we thoroughly investigate such influences by establishing a large-scale eye-tracking database of Multiple-face Video in Visual-Audio condition (MVVA). Inspired by the findings of our investigation, we propose a novel multi-modal video saliency model consisting of three branches: visual, audio and face. The visual branch takes the RGB frames as the input and encodes them into visual feature maps. The audio and face branches encode the audio signal and multiple cropped faces, respectively. A fusion module is introduced to integrate the information from three modalities, and to generate the final saliency map. Experimental results show that the proposed method outperforms 11 state-of-the-art saliency prediction works. It performs closer to human multi-modal attention.

Keywords: Visual-audio, Saliency prediction, Multiple-face video.

1 Introduction

Saliency prediction [2] is an effective way to model the deployment of possible attention on visual inputs in biological vision system. In the recent years, a surge of interest in video saliency prediction has emerged, partly because of a large number of its applications in various areas. Besides, it can be also found that most videos over the Internet contain faces, as shown in Fig. 1(a). In particular,

* Yufan Liu, Bing Li, Weiming Hu are with National Laboratory of Pattern Recognition, Institution of Automation, Chinese Academy of Sciences (CASIA), University of Chinese Academy of Sciences (UCAS) and CAS Center for Excellence in Brain Science and Intelligence Technology (CEBSIT).
† Equal contribution.
‡ Corresponding authors: Mai Xu (maixu@buaa.edu.cn), Bing Li (bli@nlpr.ia.ac.cn).

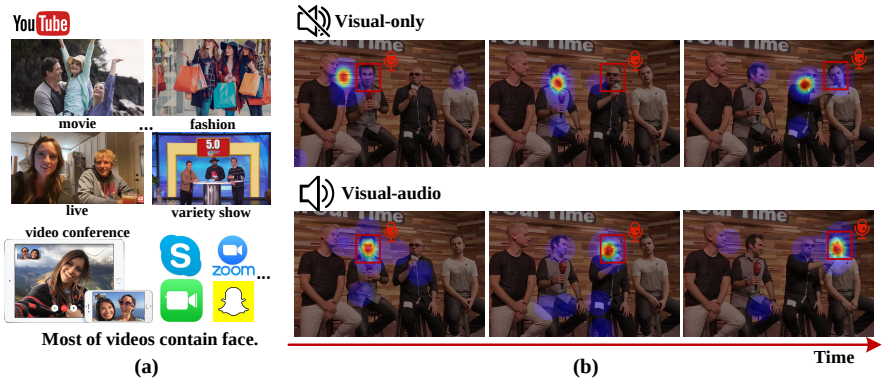

Fig. 1: (a) Thumbnails of various videos over the Internet. Most contain faces. (b) Example of visual attention on a multiple-face video. Four persons are speaking in a sequence from the left to the right. The first row ("visual-only") represents the condition when subjects view only mute frames. The second row ("visual-audio") shows the condition when both visual and audio information is present.

video conference applications (e.g., Skype and Zoom) have become popular recently, in which almost every frame has human faces. It has been reported [35] that Zoom Video Communications achieved over 39 billion annualized meeting minutes in 2018. Thus, it is necessary to predict saliency on multiple-face videos, since saliency can be used as attention cues for the content based applications, including perceptual video coding [30], quality assessment [17] and panoramic video processing [31].

Most of the video saliency works focus on the visual information and few works have taken auditory information into account. Previous works barely mention soundtracks or explicitly discard them during the eye-tracking experiments. In practice, videos are always played with sound and the world we live in always contains multi-modal information. Human attention is driven by several factors. Two most important ones include visual and auditory cues. As shown in Fig. 1(b), humans focus on different regions in visual-only condition vs. visual-audio condition. They fixate at the salient face and transit to another faster when sound is available. Without sound, people can only rely on visual cue (e.g. motion) to locate the speaking person, leading to slower attention transition. Thus, only considering visual information is not enough to predict where people look.

To address the above shortcomings, we first create a large-scale eye-tracking database dubbed Multiple-face Videos in Visual-Audio condition (MVVA). It includes fixations of 34 subjects viewing 300 videos with diverse content. To the best of our knowledge, so far this is the largest dataset of its kind. During the eye-tracking experiment, both video and audio have been presented to the viewers. Through analysis on our database, we find that faces indeed explain the majority of fixations. We further confirm that audio influences the fixation distribution

Table 1: The information of the existing visual-audio eye-tracking databases.

Database	Video Num.	Resolution	Duration	Subject	Details
Coutrot I [5]	60	720×576	10-24.8sec	20 (10 per auditory condition)	French; Scenes: one MO (Moving Object), several MO, conversation and landscapes.
Coutrot II [7]	15	720×576	12-30sec	72 (18 per auditory condition)	French; Scenes: Conversation.
Coutrot III [8]	15	1232×504	20-80s	40	English; Scenes: 4 persons meeting.
Pierre *et al.* [19]	148 (from Coutrot dbs & Hollywood)	$\leq 1232 \times 504$	0.9-35s	averaged 44 each experiment	Scenes: MO, conversation and landscapes.
Ours	300	$\geq 1280 \times 720$	10-30s	34	Chinese & English; 6 kinds of scenes

on faces and attention transition across faces. In particular, human attention in visual-audio condition significantly differs from visual-only condition, when turn-taking takes place. Inspired by these findings, we propose a novel multi-modal network to predict fixations on videos in the visual-audio condition. Our work takes faces, global visual content and audio information into consideration. It consists of three branches, namely visual, audio and face branches, to process these information respectively. Specifically, the visual branch constructs a two-stream architecture to model spatial-temporal visual saliency representation. Without other information, the output of the visual branch can be seen as the saliency map under the visual-only condition. The audio branch encodes the 1D audio signal into a 2D feature map sequence. Additionally, the face branch processes multiple cropped faces and explore the relationship between them. And then it generates a face saliency map. After that, a fusion module is introduced to integrate the three modalities, and to generate the final saliency map. We study the impact of each of these cues individually.

To summarize, our main contributions include:

- We introduce a large-scale eye-tracking database including multiple-face videos with sound, to facilitate the research on visual-audio saliency prediction.
- We present a thorough analysis on our database and study how human attention is affected by multiple factors including face and sound.
- We propose a novel multi-modal network, which fuses visual, face and audio information to obtain effective features for accurate saliency prediction.

2 Related Work

Visual saliency prediction. Visual saliency models have been widely developed to predict where people look in images [13,33,21,27,4] or videos [12,1,18,15,28,20,32]. Recently, DNNs have achieved a great success in visual saliency prediction. Over images, some deep saliency models [13,27] use multi-scale visual information to predict saliency. Over videos, most works [15,28,18] combine a CNN and an L-STM to learn spatial and temporal visual features. Bak *et al.* [1] proposed a two-stream CNN architecture. RGB frames and optical flow sequences were fed to the two streams. Zanca *et al.* [32] leveraged various visual features, such as face and motion, to predict the fixation scanpath. Recently, some works have

focused on predicting saliency over multiple-face videos. Liu *et al.* [18] presented an architecture which combined a CNN and a multiple-stream LSTM to learn face features. None of the methods above take audio modality into account. In contrast, our approach utilizes both audio and video modalities.

Visual-audio saliency prediction. Only a few methods take the auditory information into account. Early saliency models adopted hand-crafted features. For instance, in [6], low-level features (e.g., luminance information) and faces are used as visual information. Audio is fed into a speaker diarization algorithm to locate the speaking person. A saliency map is then generated by integrating the two modalities. [8] improved this method by taking the body into consideration. These methods rely heavily on the detection algorithms, which limits their performance and usability. Recent works tend to make use of learning-based methods. Tsiami *et al.* [26] combined a visual saliency model [14] and an audio saliency model [16]. But it only considers the scenario that a simple stimuli moving in clustered images. More recently, [25] used a two-stream 3D-CNN to encode visual and audio information into feature vectors, which are then concatenated to learn the final prediction.

Visual-audio databases. Few datasets have been collected for studying visual-audio attention as shown in Tab. 1. They have three main drawbacks. Firstly, they usually have a small scale. The number of videos in these datasets are typically under 100. Secondly, they contain only one or a few scenes. For example, Coutrot II [7] and Coutrot III [8] only consider eye-tracking events in a specific scene. Thirdly, their videos have low resolution. Coutrot I [5] and Coutrot II [7] contain videos with a 720 x 576 resolution. Consequently, the existing visual-audio saliency prediction methods are designed under specific conditions (e.g. under a certain scene or a low resolution). The efficiency and generalization of these models need further verification. Driven by these motivations, here we propose a dataset of 300 videos with the resolution of at least 1280 x 720 over 6 different scenes. Further, we analyze our dataset to reveal the impact of audio on human attention, and give some inspirations for saliency prediction.

3 The Proposed Dataset

In this section, we introduce a large-scale eye-tracking database called Multiple-face Video in Visual-Audio condition (MVVA). The proposed dataset contains eye-tracking fixations when both audio and video were presented. To the best of our knowledge, our dataset is the first public eye-tracking database that has multiple-face videos with audio. In addition to saliency, it can be used in other research areas such as sound localization, since the faces of speakers are manually marked in our dataset. Our dataset is publicly available in *https://github.com/MinglangQiao/MVVA-Database*.

3.1 Data collection

Stimuli. A total number of 300 videos with 146,529 frames, containing both images and audio, were collected. Among them, 143 videos were selected from

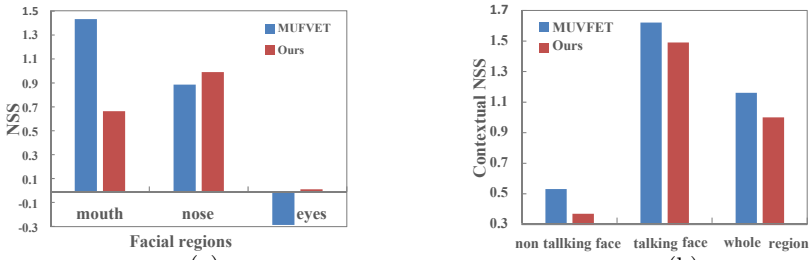

Fig. 2: (a) NSS of saliency on different facial landmarks in visual-only (MUVFET)/visual-audio (Ours) conditions. (b) Contextual NSS of optical flow maps over different face regions.

MUFVET [18] and other 157 videos were selected from YouTube, with the criterion that the videos should contain obvious faces and audio. All of them were encoded by H.264 with duration varying from 10 to 30 seconds. Note that these videos are either indoor or outdoor scenes, and can be classified into 6 categories: TV play/movie, interview, video conference, variety show, music and group discussion. The audio content covers different scenarios including quiet scenes (e.g., news broadcasting) and noisy scenes (e.g., interview at subway).

Apparatus. For monitoring the binocular eye movements, an eye tracker, EyeLink 1000 Plus [24], was used in our experiment. EyeLink1000 Plus is an integrated eye tracker with a 23.8" TFT monitor at screen resolution of 1280x720. During the experiment, EyeLink1000 Plus captured gaze data at 500 Hz. According to [24], the gaze accuracy can reach 0.25-0.5 visual degrees in the head free-to-move mode. For more details on EyeLink1000 Plus, see [24].

Participants. 34 participants (21 males and 13 females), aging from 20 to 54 (24 in average), were recruited to participate in the eye-tracking experiment. All participants had normal or corrected-to-normal vision. It is worth pointing out that only subjects who passed the eye tracking calibration were quantified for the experiment. As a result, 34 subjects (out of 39) were selected in our experiment.

Procedure. During the eye tracking experiment, all subjects were required to sit on a comfortable chair with the viewing distance of $\sim 55cm$ from the screen. Before viewing the videos, each subject was required to perform a 9-point calibration for the eye tracker. Afterwards, videos were shown in a random order and subjects were asked to view them freely. Note that the audio and video stimuli were presented simultaneously during the experiment. In order to avoid eye fatigue, the 300 videos were equally divided into 6 sessions, and there was a 5-minute rest after viewing each session. Besides, a 5-second blank period with a black screen was inserted between each two successive videos for a short break. In total we collected 5,013,980 fixations over all 34 subjects and the 300 videos.

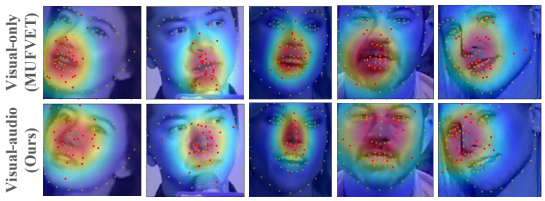

Fig. 3: Examples of saliency maps in visual-only (the first row) and visual-audio condition (the second row). The red dots are fixation points, and the yellow dots are facial landmarks.

3.2 Database analysis

Here, we thoroughly analyze our data. To annotate faces and face landmarks in video frame, we used [34] and [22], respectively, and then corrected the predictions manually. The talking/non-talking faces are manually annotated.

Finding 1: Audio influences the fixation distribution on faces. With the presence of audio, fixation distribution is different from that of visual-only scenario. First, we find that the face saliency distribution in visual-audio condition is slightly more dispersed than that in visual-only condition. We compute the averaged entropy and dispersion [19,9] of each face saliency map, and obtain 10.58 and 44.06 on our MVVA (visual-audio condition), larger than 10.16 and 39.34 of MUVFET (visual-only condition). It may be because people need to focus on mouth to identify the talking face without audio, but do not need that when audio is available. Second, as shown in Fig. 3, in the visual-audio condition, human attention tends to fixate at the center of the face (i.e., near the nose), while people tend to focus on mouth in the visual-only condition. We calculate the Normalized Scanpath Saliency (NSS) between saliency map and different facial landmarks to quantify the correlation between salient regions and facial regions in Fig. 2(a). It depicts that saliency maps in our database have the highest NSS values on nose, while on MUFVET the salient region is on mouth. This may be because people do not need to concentrate on the mouth motion, when they can clearly hear the sound. Third, attention transits from mouth/nose to eyes when face becomes larger. We compute NSS of saliency map on facial landmarks, and calculate the Pearson correlation coefficient between the NSS and the normalized face size. We find that the Pearson correlation coefficients between face size and NSS on {eyes, mouth, nose} in order are {0.29, -0.44, -0.12} in our dataset, and {0.54, -0.49, 0.14} in MUFVET. Positive correlation between face size and NSS on eyes reflects more attention on eyes when subjects are viewing larger faces.

Finding 2: In the turn-taking scenes, the transition of fixations across faces is largely influenced by audio. Fig. 4(a) shows an example of attention transition in the turn-taking scenes. It can be observed that human fixations transit and follow the talking face faster in the visual-audio condition than that in the visual-only condition. Fig. 1 also shows the similar observation. For quantitative analysis, we compare the attention transition time in visual-audio and visual-only conditions. We define the attention transition time by the average number of frames that fixations transit to the talking face, when turn-taking happens. Here, F_{va} and F_{vo} denote the attention transition time in MVVA

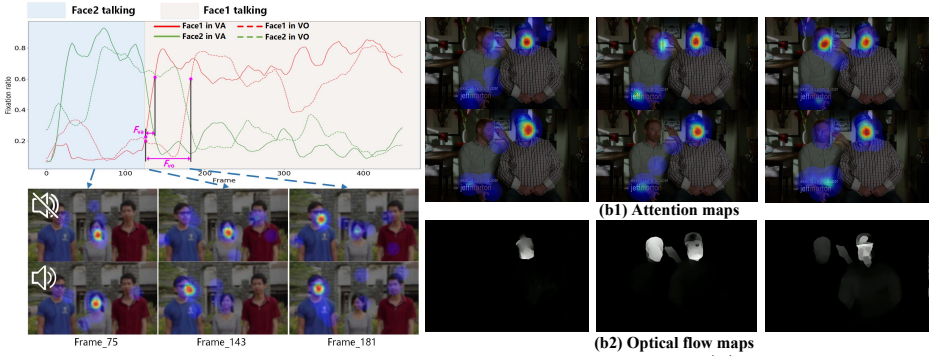

Fig. 4: (a) An example of attention transition in Visual-Only (VO, the first row of heat maps) and Visual-Audio condition (VA, the second row of heat maps). (b1) One video example showing the saliency difference between visual-only condition (the first row) and visual-audio condition (the second row). The person at the right is talking while the other is turning his head. (b2) The corresponding optical flow maps of each frame.

(visual-audio condition) and MUVFET (visual-only condition), respectively. The results of F_{va} and F_{vo} are 30 and 24 frames. Thus, the attention transition time in visual-audio condition is shorter than that in visual-only condition by 25%. From the above results, we can conclude that the fixations transit across faces are largely influenced by audio.

Finding 3: Human attention is more influenced by motion in the absence of audio. It is intuitive that people are guided by the visual cues (e.g., motion) more in the visual-only condition, compared to the visual-audio condition. This is because people can only rely on the visual cues to figure out what is going on in the video under the visual-only condition. For instance, in Fig. 4(b), in visual-only condition attention is mostly attracted to the person on the left who is turning his head, while in the visual-audio condition, subjects concentrate on the right speaking person. To quantify the relationship between motion and saliency, we computed the contextual NSS [25] of the optical flow maps on fixations. Fig. 2(b) illustrates that human attention correlates more with motion in the visual-only condition.

4 The Proposed Method

According to the findings above, visual information, audio and faces are all important factors that influence human attention. In this section, we introduce our multi-modal saliency method that utilizes these information for predicting fixations over multiple-face videos. Fig. 5 summarizes the overall framework of

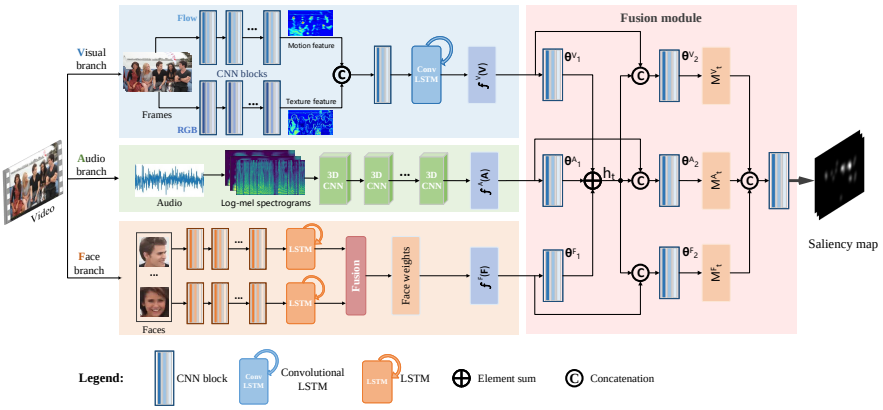

Fig. 5: Overall framework of the proposed method.

the proposed method. A three-branch neural network is used to integrate multiple information cues and to generate a saliency map. Particularly, a video segment $Video = \{\mathbf{V}, \mathbf{A}, \mathbf{F}\}$, comprising visual frames $\mathbf{V} = \{V_t\}_{t=1}^{T}$, audio signals $\mathbf{A} = \{A_t\}_{t=1}^{T}$ and faces $\mathbf{F} = \{F_t\}_{t=1}^{T}$, is first fed into our multi-modal neural network. Each component of the video segment is conveyed to the corresponding branch of the network. The predicted saliency maps $\mathbf{S} = \{S_t\}_{t=1}^{T}$ are then computed as:

$$\mathbf{S} = f(\mathbf{V}, \mathbf{A}, \mathbf{F}) = \Phi(f^V(\mathbf{V}), f^A(\mathbf{A}), f^F(\mathbf{F})), \tag{1}$$

where $f(\cdot)$ is the proposed model, and $f^V(\cdot), f^A(\cdot), f^F(\cdot)$ are the three branches for visual, audio and face cues, respectively. Besides, $\Phi(\cdot)$ is the fusion module to integrate the three modalities and to generate the final saliency maps.

4.1 Architecture

Visual branch. Fig. 5 shows visual branch constructs a two-stream CNN & convolutional LSTM architecture to model spatial-temporal visual representation. In detail, on the one hand, the frames $\mathbf{V}$ are fed to an RGB sub-branch to obtain the features of texture. On the other hand, frames are fed to a flow sub-branch to get the features of motion. Note that the flow sub-branch is initialized by FlowNet [10] so that it can obtain motion-oriented features. Then, these extracted features are concatenated (denoted as $\mathrm{C}(\cdot)$) and are fed to a two-layer convolutional LSTM [29], which is leveraged to process spatial-temporal information. After that, feature maps $f^V(\mathbf{V})$ are obtained as:

$$f^V(\mathbf{V}) = \mathrm{LSTM}(\mathrm{C}(g_1(\mathbf{V}), g_2(\mathbf{V}))). \tag{2}$$

Note that $g_1(\cdot)$ represents the RGB sub-branch, consisting of four CNN blocks of VGG-16 [23]. And $g_2(\cdot)$ denotes the flow sub-branch, which comprises three CNN blocks and one deconvolutional layer of FlowNet.

Audio branch. In audio branch, a frequency domain based 3D-CNN is designed to convolute 1D audio signal by converting it to 2D spectrum. As such, the spectrum can be better integrated with 2D image features. In detail, the audio signal is first re-sampled to 22kHz and is then transformed to log-mel spectrogram using Short-Time Fourier Transform (STFT) and mel-mapping [11], with a hop length of 512. To be consistent with the visual frame, the log-mel spectrogram is converted into a sequence of successive overlapping frames, and is cropped in a $(-230, 230]$ ms window. After that, 4-layer 3D-CNNs $g_{3d}(\cdot)$ are embedded to encode the log-mel spectrogram sequence and to obtain the audio feature maps:

$$f^A(\mathbf{A}) = g_{3d}(\text{STFT}(\mathbf{A})). \tag{3}$$

Face branch. In face branch, a dynamic multi-stream spatial-temporal LSTM model is designed for exploring relationship between multi-faces with features interacting with each other. Fig. 6(a) gives a detailed illustration of the face branch. Firstly, given a sequence of video frames, the MTCNN face detector [34] is leveraged to detect and crop faces. Secondly, N cropped faces are fed into N parameter-shared sub-branches containing an 13-layer CNNs and a 2-layer LSTM, and are transformed to N feature vectors. After that, these features are fused by the fusion part of face branch, which helps face features to capture the correlation and competition with each other. Hence, each face sub-branch perceives the sufficient information and we can obtain N face saliency weights: $\mathbf{w}_1 = \{w_{1,t}\}_{t=1}^{T}, \mathbf{w}_2, ..., \mathbf{w}_N$. Larger weight for a face means that it is more salient. Finally, we calculate the face feature map $f^F(F_t)$ at the t-th frame as follows,

$$f^F(F_t) = \sum_{n=1}^{N} w_{n,t} \cdot \mathcal{N}_{n,t}. \tag{4}$$

Here, we follow [18] to regard saliency on the n-th face as a Gaussian distribution $\mathcal{N}_{n,t}(\mu_{n,t}, \Sigma_{n,t})$[1].

The parameter-sharing architecture can process videos with different face numbers. As shown in Fig. 6(b), a new CNN-LSTM stream is instantiated when there is a new face appearing in the video. To be specific, we use PyTorch to instantiate CNN-LSTM streams with different number at each iteration.

In the training process, firstly we pre-train the face branch. The fixation proportion of the n-th face to all faces at frame t (denoted as $w_{n,t}$) is taken as the Ground Truth (GT) weight to supervise the predicted face saliency weight (denoted as $\hat{w}_{n,t}$). Hence, the optimization can be formulated as

$$\min \sum_{t=1}^{T} \sum_{n=1}^{N} ||\hat{w}_{n,t} - w_{n,t}||_2^2, \quad s.t. \sum_{n=1}^{N} \hat{w}_{n,t} = 1. \tag{5}$$

Fusion. After encoding each video modality to feature maps, the proposed model integrates visual, audio and face feature maps together to learn a joint representation. We propose a fusion module depicted in Fig. 5, instead of direct concatenation. Given visual, audio and face feature maps $\{f^F(V_t), f^F(A_t), f^F(F_t)\}$, the

[1] $\mathcal{N}_{n,t}(\mathbf{x}) = \exp\{-\frac{1}{2}(\mathbf{x} - \mu_{n,t})^T \Sigma_{n,t}^{-1}(\mathbf{x} - \mu_{n,t})\}$

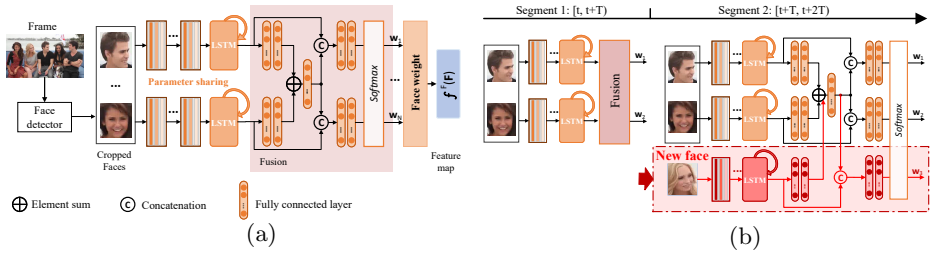

Fig. 6: (a) Structure of face branch. (b) An example of face branch processing variant face numbers.

fusion module performs the computations below:

$$\begin{aligned} \mathrm{M}_t^V &= \Theta_2^V * \mathrm{C}(h_t, f^V(V_t)), \\ \mathrm{M}_t^A &= \Theta_2^A * \mathrm{C}(h_t, f^A(A_t)), \\ \mathrm{M}_t^F &= \Theta_2^F * \mathrm{C}(h_t, f^F(F_t)), \\ s.t. \quad h_t &= \Theta_1^V * f^F(V_t) + \Theta_1^A * f^F(A_t) + \Theta_1^F * f^F(F_t). \end{aligned} \tag{6}$$

Note that the Θs are the parameters of different CNN blocks, which align multi-modal features with different scales and receptive fields (e.g., visual branch outputs global features while face branch outputs local features). And '$*$' denotes convolution operator and $\mathrm{C}(\cdot)$ is the concatenation operation. With help of the fusion module, the three branches can share information and preserve original characteristics of themselves.

4.2 Optimization

To train and optimize the proposed multi-modal network, we use the GT fixation map $\mathbf{G}$, obtained from the fixation density map, to supervise the predicted saliency map $\mathbf{S}$. The loss function is the Kullback-Leibler (KL) divergence between the two maps,

$$\mathbf{L} = \sum_{t=1}^{T} KL(G_t||S_t) = \sum_{t=1}^{T} \sum_{i \in \mathbf{I}} G_t(i) \log \frac{G_t(i)}{S_t(i)}, \tag{7}$$

in which i denotes a position in the 2D saliency map. Note that KL divergence is chosen because Huang *et al.* [13] have proven that the KL divergence is more effective than other metrics in training DNNs for predicting saliency. To make the convergence speed faster, we pre-train the three branches. In particular, the visual and face branches are pre-trained on MVVA separately. For the visual branch, the RGB sub-branch is initialized with VGG parameters on ImageNet, while the Flow sub-branch is initialized with FlowNet parameters. The face branch is also initialized with VGG. Then, the audio branch is pre-trained jointly with the visual branch, since only audio cannot locate salient faces.

Table 2: Accuracy of saliency prediction by our method and 11 competing methods over different datasets.

		Ours	TASED	SAM_res	SAM_vgg	Liu	ACLNet	DeepVS	SalGAN	Coutrot	SALICON	OBDL	BMS	G-Eymol
MVVA	AUC	**0.905**	0.905	0.897	0.896	0.893	0.889	0.890	0.891	0.869	0.866	0.786	0.765	0.615
	NSS	**3.976**	3.319	3.495	3.466	3.279	3.437	3.270	2.650	2.604	2.523	1.342	0.936	0.551
	CC	**0.722**	0.653	0.634	0.634	0.625	0.639	0.615	0.539	0.509	0.477	0.273	0.193	0.125
	KL	**0.823**	0.970	1.004	1.012	1.098	1.044	1.117	1.234	1.557	1.447	1.995	2.051	4.253
Coutrot II [7]	AUC	**0.922**	0.877	0.905	0.849	0.908	0.848	0.896	0.900	0.883	0.865	0.723	0.751	0.698
	NSS	**3.568**	2.731	3.446	3.306	2.833	3.127	3.058	2.286	3.033	2.408	0.730	0.739	0.884
	CC	**0.639**	0.545	0.607	0.593	0.585	0.521	0.556	0.553	0.606	0.433	0.181	0.153	0.162
	KL	**0.915**	1.271	1.031	1.093	1.035	1.357	1.209	1.717	1.428	1.514	2.228	2.073	2.932

5 Experiments and Results

5.1 Settings

In our experiment, 300 videos in our MVVA are randomly divided into training (240 videos) and test (60 videos) sets. Specifically, for the visual branch, RGB frames are resized to 256x256. To train the convolutional LSTM, we temporally segment 240 training videos into 9,806 clips, all of which have $T = 12$ frames. For the audio branch, we use the 16-frame segmented log-mel spectrograms which are also resized to 256x256. For the face branch, the resolution of N input faces is 128x128. The parameters of the proposed network are updated by using the Stochastic Gradient Descent (SGD) algorithm with Adam optimizer. The initial learning rate is set to be 1e-4.

To evaluate our method, we adopt four metrics: Area Under the receiver operating Characteristic curve (AUC), NSS, Correlation Coefficient (CC), and KL divergence [2]. Note that the larger values for AUC, NSS or CC indicate more accurate saliency prediction. The opposite holds for the KL divergence. Please see [3] for more details on these metrics. All experiments are conducted on a computer with Intel(R) Core(TM) i7-8700 CPU@3.20 GHz, 62.8 GB RAM and 2 Nvidia GeForce GTX 1080 Ti GPUs.

5.2 Performance Comparison

We compare the performance of our multi-modal method with 11 state-of-the-art saliency prediction methods, including TASED [20], SAM [4], Liu [18], ACLNet [28], DeepVS [15], SalGAN [21], SALICON [13], Coutrot [8], OBDL [12], BMS [33] and G-Eymol [32]. Among them, SalGAN, SALICON, SAM and BMS are state-of-the-art saliency prediction methods over images, and others are for videos. SAM has two versions, including SAM_res with ResNet backbone and SAM_vgg with VGGNet backbone. Note that Coutrot and Liu focus on multiple-face videos. In Coutrot, static saliency map, dynamic saliency map, speaker map and center bias map are weighted with estimated weights, and merged into the final saliency map. To eliminate the influence of the feature extraction algorithm (e.g., face/speaking detection), we re-implement Coutrot *et al.* method with manual annotated features and treat the performance as the

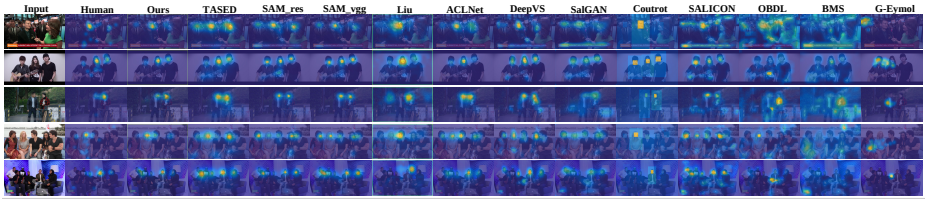

Fig. 7: Saliency maps of 5 videos randomly selected from the test set of our eye-tracking database.

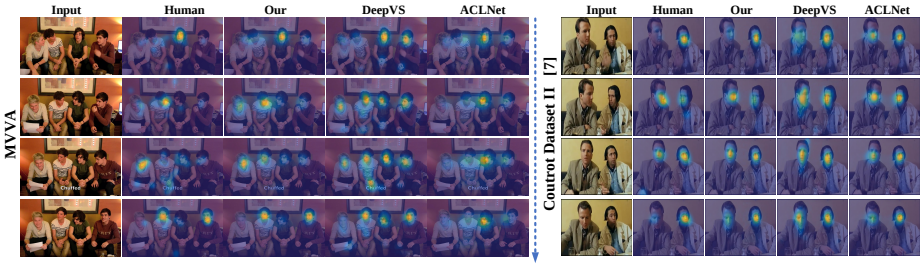

Fig. 8: Saliency maps for different frames of two video sequences, selected from our MVVA and Coutrot II [7].

upper bound of Coutrot *et al.* Liu is the latest DL based method for multiple-face videos, but it ignores the audio information. Besides, face is also considered in G-Eymol as a semantic-based feature. To effectively assess the power of our method, we test it on different databases as follows.

Evaluation on our dataset. Tab. 2 presents AUC, NSS, CC and KL divergence for the proposed method versus 11 competing methods. Scores are averaged over 60 test videos in our eye-tracking database. As shown in this table, the proposed method performs significantly better than all other methods over all 4 metrics. Specifically, compared with the best competing result, our method achieves over 0.481, 0.069 and 0.147 improvements in NSS, CC and KL, respectively. The main reasons for this result are: 1) Most of state-of-the-art methods do not consider audio information, while our method does utilize audio cue for saliency prediction, 2) The face temporal subnet of our method learns detailed face features to predict salient faces, and 3) Our fusion module effectively integrates the multi-modal information.

Next, we compare models qualitatively. Fig. 7 demonstrates saliency maps over 5 randomly selected videos in the test set, predicted by the proposed method and 11 other methods. As is shown, our method is capable of locating the salient faces. Its prediction is much closer to the GT. Besides, the proposed method shows excellent performance on predicting attention transition, as depicted in Fig. 8. In contrast, most of the other methods fail to accurately predict the

Table 3: Performance of different modules in our model.

	Models	CC	KL	NSS	AUC
Different modules	Visual (RGB only)	0.527	1.324	2.728	0.860
	Visual (Flow only)	0.510	1.354	2.631	0.869
	Visual (RGB+flow)	0.632	1.043	3.358	0.893
	Visual (RGB+flow+LSTM)	0.671	0.971	3.548	0.896
	Visual+audio	0.712	0.843	3.838	0.907
	Face only	0.569	1.292	2.766	0.872
	Face+audio	0.609	1.116	3.211	0.878
	Visual+audio+face	**0.722**	**0.823**	**3.976**	**0.905**

regions that attract human attention, perhaps because these methods do not consider extra information such as sound and face.

Evaluation on generalization ability. To evaluate the generalization capability of the proposed method, we further evaluate our method and 11 other methods on the Coutrot II database [7]. Tab. 2 compares the average AUC, NSS, CC and KL scores. As shown in this table, the proposed method again outperforms all the competing methods. In particular, there are at least 0.032 and 0.116 improvements in CC and KL, respectively. Such improvements are comparable to those in our MVVA. Qualitative results, shown in Fig. 8, shows the proposed method predicts attention transition accurately, while other methods miss salient faces. These results demonstrate the generalization capability of our method in video saliency prediction.

5.3 Ablation Analysis

Here, we thoroughly analyze the effectiveness of each module in our method.

Visual branch. Visual branch uses basic visual information, i.e., texture, motion and temporal cues, to predict saliency. We evaluate the visual branch of the proposed network and report the results in Tab. 3. It shows that visual branch reaches to CC of 0.632 and KL of 1.043, which is better than many methods and comparable with the best competing method TASED. When adding convolutional LSTM to fuse the temporal cues, the performance reaches to 0.671 in CC and 0.971 in KL. Hence, the entire visual branch and its components are all useful to saliency prediction. Moreover, as shown in Tab. 3, combination of face and audio results in lower performance than combining all cues (i.e., the whole network, visual+audio+face) by a large margin. It further manifests the effectiveness of visual branch. We add visual branch, because there are still some other regions drawing attention, besides faces.

Audio branch. Besides visual branch, we add audio branch to the framework. With the help of the audio branch, the visual-audio model achieves 0.712 in CC and 0.843 in KL, much better than the visual branch. In addition, the combination of face and audio branches improves the performance of the single face branch, by 0.040 in CC and 0.176 in KL. Thus, these results manifest the contribution of audio information and the effectiveness of the proposed audio branch.

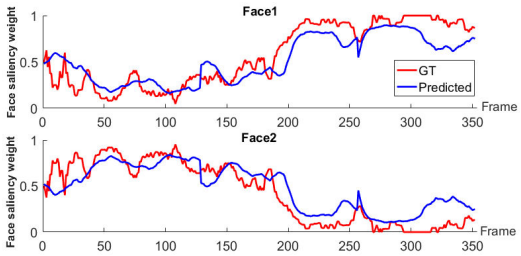

Fig. 9: Face saliency weights across frames for a randomly selected video.

Face branch. Finally, the face branch is added to complete the whole network. From Tab. 3, CC of 0.722 and KL of 0.823 are reached, after combining face branch with visual-audio model. It is worth mentioning that the single face branch can only achieve a fair performance, which is inferior to other combinations. Hence, single face branch cannot reach the best accuracy, even most attention is attracted by faces. In addition, since the face branch aims at predicting saliency weight of faces across the video frames, we plot the face saliency weights of the proposed face branch and GT in Fig. 9. In this figure, the curve of the face branch fits close to the curve of GT. It can be concluded that the face branch accurately predicts the salient face and further enhances the performance of the proposed model.

In summary, the ablation analysis manifests the necessity of different cues for saliency prediction, and verifies the effectiveness of each part in our model. More details can be found in the supplementary document.

6 Conclusion

In this paper, we explored how audio influences human attention in multiple-face videos. Various findings have been verified by the statistical analysis on our new eye-tracking database. To predict multiple-face video saliency, we presented a novel multi-modal network consisting of visual, audio and face branches. The three branches encode visual frames, audio spectrograms and faces into feature maps, respectively. A fusion module was designed to integrate the three modalities, and to generate the final saliency map. Finally, experimental results shown that our method outperforms 11 state-of-the-art methods over several datasets.

Acknowledgement. This work is supported by Beijing Natural Science Foundation (Grant No. L172051, JQ18018), the Natural Science Foundation of China (Grant No. 61902401, 61972071, 61751212, 61721004, 61876013, 61922009, 61573037 and U1803119), the NSFC-general technology collaborative Fund for basic research (Grant No. U1636218, U1936204), CAS Key Research Program of Frontier Sciences (Grant No. QYZDJ-SSW-JSC040), CAS External cooperation key project, and NSF of Guangdong (No. 2018B030311046). Bing Li is also supported by CAS Youth Innovation Promotion Association.

References

1. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Transactions on Multimedia **20**(7), 1688–1698 (2017)
2. Borji, A.: Saliency prediction in the deep learning era: An empirical investigation. arXiv preprint arXiv:1810.03716 (2018)
3. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
4. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing **27**(10), 5142–5154 (2018)
5. Coutrot, A., Guyader, N.: Toward the introduction of auditory information in dynamic visual attention models. In: International Workshop on Image Analysis for Multimedia Interactive Services. pp. 1–4. IEEE (2013)
6. Coutrot, A., Guyader, N.: An audiovisual attention model for natural conversation scenes. In: IEEE International Conference on Image Processing. pp. 1100–1104. IEEE (2014)
7. Coutrot, A., Guyader, N.: How saliency, faces, and sound influence gaze in dynamic social scenes. Journal of Vision **14**(8), 5–5 (2014)
8. Coutrot, A., Guyader, N.: An efficient audiovisual saliency model to predict eye positions when looking at conversations. In: European Signal Processing Conference. pp. 1531–1535. IEEE (2015)
9. Coutrot, A., Guyader, N., Ionescu, G., Caplier, A.: Influence of soundtrack on eye movements during video exploration. Journal of Eye Movement Research **5**(4), 2 (2012)
10. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: IEEE International Conference on Computer Vision. pp. 2758–2766 (2015)
11. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 131–135. IEEE (2017)
12. Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I.V., Shan, Y.: How many bits does it take for a stimulus to be salient? In: IEEE Conference on Computer Vision and Pattern (2015)
13. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: IEEE International Conference on Computer Vision (2015)
14. 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 (11), 1254–1259 (1998)
15. Jiang, L., Xu, M., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: European Conference on Computer Vision. pp. 602–617 (2018)
16. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology **15**(21), 1943–1947 (2005)

17. Li, C., Xu, M., Du, X., Wang, Z.: Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In: ACM International Conference on Multimedia. pp. 932–940 (2018)
18. Liu, Y., Zhang, S., Xu, M., He, X.: Predicting salient face in multiple-face videos. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4420–4428 (2017)
19. Marighetto, P., Coutrot, A., Riche, N., Guyader, N., Mancas, M., Gosselin, B., Laganiere, R.: Audio-visual attention: Eye-tracking dataset and analysis toolbox. In: IEEE International Conference on Image Processing. pp. 1802–1806. IEEE (2017)
20. Min, K., Corso, J.J.: Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection (2019)
21. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
22. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1685–1692 (2014)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. SR-Research: Eyelink 1000 plus, https://www.sr-research.com/products/eyelink-1000-plus/
25. Tavakoli, H.R., Borji, A., Rahtu, E., Kannala, J.: Dave: A deep audio-visual embedding for dynamic saliency prediction. arXiv preprint arXiv:1905.10693 (2019)
26. Tsiami, A., Katsamanis, A., Maragos, P., Vatakis, A.: Towards a behaviorally-validated computational audiovisual saliency model. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2847–2851. IEEE (2016)
27. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**(5), 2368–2378 (2017)
28. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: IEEE conference on Computer Vision and Pattern Recognition. pp. 4894–4903 (2018)
29. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information ProcessingSsystems. pp. 802–810 (2015)
30. Xu, M., Liu, Y., Hu, R., He, F.: Find who to look at: turning from action to saliency. IEEE Transactions on Image Processing **27**(9), 4529–4544 (2018)
31. Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., Wang, Z.: Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2693–2708 (2019)
32. Zanca, D., Melacci, S., Gori, M.: Gravitational laws of focus of attention. IEEE (2019)
33. Zhang, J., Sclaroff, S.: Exploiting surroundedness for saliency detection: a boolean map approach. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 889–902 (2016)
34. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
35. Zoom: Zoom announces lineup of global technology and thought leaders for zoomtopia 2018, https://blog.zoom.us/wordpress/2018/07/11/zoom-announces-lineup-of-global-technology-and-thought-leaders-for-zoomtopia-2018/