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

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.

In this supplementary document, we present more details about the proposed dataset, method and experiments.

1 More details about MVVA

As is shown in Figure 1 and Table 1, the multiple-face videos in our large-scale dataset are at diverse scenarios, and can be categorized into 6 classes, including TV play/movie, interview, video conference, variety show, music and group discussion. In addition, the audio content covers different scenarios including quiet scenes and noisy scenes, as reported in Table 2. In the noisy scenes, the background sounds contain laughter, street, music, applause, crowd and noise.

Moreover, we attach several attention map videos in this supplementary material. In particular, "fig1_video.mp4", "fig4a_video.mp4" and "fig4b_video.mp4" correspond to Fig.1, Fig.4a and Fig.4b in this paper, respectively. In "fig1_video.mp4" and "fig4a_video.mp4", attention transits from one face to another faster when sound is available. It can be also seen that human attention tends to fixate at the center of the face in the visual-audio condition, while people tend to focus on mouth in the visual-only condition. In "fig4b_video.mp4", we can find that human attention is more influenced by motion in the absence of audio. The left person who is turning his head attracts more attention in visual-only condition, while in the visual-audio condition, subjects mainly concentrate on the speaking person on the right side.

Table 1. Video categories in our database.

Category	TV play/movie	interview	video conference	TV show	music/talk show	group discussion	**overall**
Number of videos	53	71	14	67	51	44	300

* 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. One example for each category of videos. From left to right, the videos belong to TV play/movie, interview, video conference, TV show, music/talk show, and group discussion.

Table 2. Audio scenes categories in our database.

Category	noisy scenes						quiet scenes	**overall**
	laughter	street	music	applause	crowd	noise		
Number of videos	34	17	72	16	46	19	96	300

2 More findings

Finding 4: Subjects are consistent in terms of where they look and concentrate on the same face.

With audio and video presented simultaneously, we find that fixations are highly consistent. We randomly and equally divide the subjects into two non-overlapping groups: group A and group B, by 20 trails. Then, the Correlation Coefficient (CC) between fixation heat maps of group A and B are calculated, which is 0.75 (standard deviation: 0.08). This implies high consistency across subjects in viewing multiple-face videos.

On the other hand, the proportion of the fixations falling into face regions is 70.0%. Among the fixations in face regions, the proportion of the fixations falling into the same face is 73.8%. Thus, it can be concluded that people tend to concentrate on faces, especially the same face.

3 More details about the proposed method

In this paper, we propose a three-branch multi-modal network to predict saliency on multiple-face videos. In detail, the configuration of the proposed network is reported in Table 3. In visual branch, the RGB frames are fed into RGB sub-branch and flow sub-branch. The output features of the two sub-branches are then concatenated and are fed to a two-layer convolutional LSTM to get the visual feature maps. In audio branch, the log-mel spectrogram sequence is encoded to audio feature maps by several 3D-CNNs. Besides, multiple cropped faces are fed into the face branch and obtain the face saliency weights. By generating face conspicuity map using Gaussian distribution following [1], the face feature maps are obtained. Finally, the fusion module is proposed to integrate the three types of feature maps and generates the saliency map.

4 More experimental results

Here, we provide some predicted saliency videos in this supplementary material. The thumbnails of these videos are depicted in Figure 2. Note that

Table 3. The configuration of the proposed network.

Visual branch						**Audio branch**			**Face branch**		
RGB sub-branch			**Flow sub-branch**								
Blocks	Height x Width x Channel	Kernels ([size, stride]xnum)	Blocks	Height x Width x Channel	Kernels ([size, stride]xnum)	Blocks	(Time x) Height x Width x Channel	Kernels ([size, stride]xnum)	Blocks	Height x Width x Channel	Kernels ([size, stride]xnum)
input	256 x 256 x 3	-	input	256 x 256 x 6	-	input	16 x 64 x 64 x 1	-	input	128 x 128 x 3	-
conv	256 x 256 x 64	[3x3, 1] x 2	flow_conv	128 x 128 x 64	[7x7, 2] x 1	conv_3d	16 x 64 x 64 x 16	[3x3x3,1] x 1	conv	256 x 256 x 64	[3x3, 1] x 2
maxpool_2d	128 x 128 x 64	[2x2, 2] x 1	flow_conv	64 x 64 x 128	[5x5, 2] x 1	maxpool_3d	8 x 32 x 32 x 16	[2x2x2,2] x 1	maxpool_2d	128 x 128 x 64	[2x2, 2] x 1
conv	128 x 128 x 128	[3x3, 1] x 2	flow_conv	32 x 32 x 256	[5x5, 2] x 1	conv_3d	8 x 32 x 32 x 32	[3x3x3,1] x 2	conv	128 x 128 x 128	[3x3, 1] x 2
maxpool	64 x 64 x 128	[2x2, 2] x 1	maxpool	16 x 16 x 256	[2x2, 2] x 1	maxpool_3d	4 x 32 x 32 x 32	[2x1x1,2x1x1] x 1	maxpool_2d	64 x 64 x 128	[2x2, 2] x 1
conv	64 x 64 x 256	[3x3, 1] x 1	deconv	32 x 32 x 128	[4x4, 2] x 1	conv_3d	4 x 32 x 32 x 64	[3x3x3,1] x 1	conv	32 x 32 x 256	[3x3, 1] x 3
maxpool_2d	32 x 32 x 256	[2x2, 2] x 1	-	-	-	maxpool_3d	2 x 32 x 32 x 64	[2x1x1,2x1x1] x 1	maxpool_2d	16 x 16 x 256	[2x2, 2] x 1
conv	32 x 32 x 512	[3x3, 1] x 1				reshape	32 x 32 x 128	-	conv	16 x 16 x 512	[3x3, 1] x 6
Combination									avgpool_3d	1 x 1 x 256	[16x16x2, 1x1x2] x1
concat	32 x 32 x 640	-							LSTM	1024	[1024] x 2
conv	32 x 32 x 256	[3x3, 1] x 1				-	-	-	FC	1024	[1024] x 1
conv	32 x 32 x 128	[3x3, 1] x 2							FC	1024	[1024] x 1
convLSTM	32 x 32 x 128	[3x3, 1] x 1							FC	1	[1] x 1
convLSTM	32 x 32 x 128	[3x3, 1] x 1							output	32 x 32 x 1	-
Fusion module											
conv	32 x 32 x 64	[3x3, 1] x 2				conv	32 x 32 x 64	[3x3, 1] x 2	conv	32 x 32 x 64	[3x3, 1] x 2
concat	32 x 32 x 1	[1x1, 1] x 1									

"fig8_video_MVVA.mp4" and "fig8_video_CoutrotII.mp4" correspond to Fig.8 in this paper. It can be seen that our predicted results are close to the GT and correctly locate the salient faces.

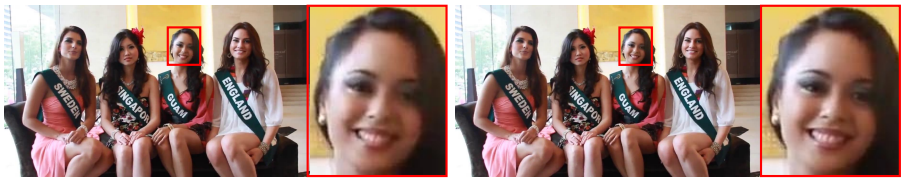
On the other hand, to further discuss the effectiveness of the fusion module, we test the model which simply fuses the three branches by concatenation operation, similar to [2]. The performance of this model is 0.705 of CC, 0.862 of KL, 3.839 of NSS and 0.908 of AUC, which is inferior to the proposed method. Hence, the proposed fusion module that integrates different modalities is necessary.

(a) (b) (c)

Fig. 2. The thumbnails of the human (GT) attention map videos and our predicted saliency map videos.

5 Applications

The proposed saliency prediction method has potential to be implemented in some tasks of perceptual video processing. For instance, more bits can be assigned to salient faces for perceptual video coding. We have applied our saliency prediction method in the compression of video conferencing, which have been put into use in some video corporation. At the same perceptual quality, it can save approximately 40% bit-rate on the compressed videos. Equivalently, the



(a) Conventional compression **(b) Perceptual compression**

Fig. 3. Subjective quality comparison. (a) and (b) are the 311-th frame compressed at 650K bit-rate by the conventional and perceptual compression, respectively. The perceptual compression is implemented using our method.

perceptual quality can be significantly improved at the same compression bit-rates. Fig. 3 shows such an example. It can be observed that the perceptual compression using our saliency prediction method yields more satisfactory quality in the salient face, compared to the conventional compression. In summary, the effect on some applications (e.g., perceptual video coding) emphasize the important role the task act.

References

1. Liu, Y., Zhang, S., Xu, M., He, X.: Predicting salient face in multiple-face videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4420–4428 (2017)
2. 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)