Towards Audio-Visual Saliency Prediction for Omnidirectional Video with Spatial Audio

Fang-Yi Chao†, Cagri Ozcinar‡, Lu Zhang†, Wassim Hamidouche†, Olivier Deforges†, Aljosa Smolic‡
†Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
‡V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland

***Abstract*—Omnidirectional videos (ODVs) with spatial audio enable viewers to perceive 360° directions of audio and visual signals during the consumption of ODVs with head-mounted displays (HMDs). By predicting salient audio-visual regions, ODV systems can be optimized to provide an immersive sensation of audio-visual stimuli with high-quality. Despite the intense recent effort for ODV saliency prediction, the current literature still does not consider the impact of auditory information in ODVs. In this work, we propose an audio-visual saliency (AVS360) model that incorporates 360° spatial-temporal visual representation and spatial auditory information in ODVs. The proposed AVS360 model is composed of two 3D residual networks (ResNets) to encode visual and audio cues. The first one is embedded with a spherical representation technique to extract 360° visual features, and the second one extracts the features of audio using the log mel-spectrogram. We emphasize sound source locations by integrating audio energy map (AEM) generated from spatial audio description (*i.e.*, ambisonics) and equator viewing behavior with equator center bias (ECB). The audio and visual features are combined and fused with AEM and ECB via attention mechanism. Our experimental results show that the AVS360 model has significant superiority over five state-of-the-art saliency models. To the best of our knowledge, it is the first work that develops the audio-visual saliency model in ODVs. The code is publicly available at: https://github.com/FannyChao/AVS360_audiovisual_saliency_360.**

***Index Terms*—Audio-visual saliency, spatial sound, ambisonics, omnidirectional video (ODV), virtual reality (VR).**

I. Introduction

Recent technical advances in head-mounted displays (HMDs) have paved the way for the present virtual reality (VR) technologies to move out of research lab environments into our daily life. From video games to narrative films, omnidirectional video (ODV) is changing how viewers interact and perceive video by providing immersive sensation within a scene with the help of HMDs. The audio-visual representation of ODV is captured with omnidirectional microphone and multi-camera systems. The audio part of ODV can be represented by spatial audio, *e.g.*, ambisonics, which is a description of a 3D spatial audio scene. It enables viewers to perceive the directions of sound corresponding to their heads positions when they are rotating heads to explore ODVs. The visual part of the ODV signal, which is a 3D spherical representation of a 360° scene around a given center, is typically stored in 2D planar formats, *e.g.*, equirectangular projection (ERP), to be compatible with the existing video technology systems. Thanks to its immersive and interactive nature, there has been increasing interest in the audio-visual aspects of ODV. For example, Rana *et al.* [1] and Morgado *et al.* [2] use texture and mono audio of ODV to predict its audio source location.

For optimizing VR systems, such as streaming [3], it is essential to understand and anticipate human behavior while watching ODVs. Recent visual attention studies for ODV have set a fundamental background for analyzing users' behavior in VR systems. David *et al.* [4] and Ozcinar *et al.* [5], for instance, reveal that visual attention in ODVs tends to be concentrated in the *equator center* of ODV. In addition, Chao *et al.* [6] show that spatial audio *sound source location* in ODVs is an important visual attention feature. They show that different audio-visual contents (*i.e.*, conversation, music, and environment) of ODVs and different audio modalities (*i.e.*, mute, mono, and ambisonics) have a different interactive effect on human visual saliency. However, the audio part of ODV is highly overlooked by existing computational ODV saliency models.

To more precisely predict visual saliency in ODVs, in this paper, we propose a new computational audio-visual saliency (AVS360) model for ODV that incorporates omnidirectional visual representation and 360° spatial audio cues. Inspired by [7], we use the two-stream structure to encode visual and audio cues separately. For visual, we consider the geometric nature of ODV by embedding spherical visual representation of ODV using Cube Padding (CP) technique [8] into a 3D ResNet. This modification can help to extract spherical spatial-temporal features in ODVs. For audio, we integrate spatial audio description into a 3D ResNet via attention mechanism to emphasize the locations of sound sources. To take the effect of audio-visual cues, we concatenate and fuse audio and visual features with 2D convolutional neural layers. Inspired by the finding in recent studies [4]–[6], [9], we also consider equator center viewing behavior and the impact of sound source locations to improve saliency prediction accuracy in ODV. For this purpose, we use equator center bias (ECB) and audio energy map (AEM) in the fusion process. To the best of our knowledge, it is the first work in the literature that explicitly tackles the problem of audio-visual saliency prediction. Our model was trained with an eye-tracking dataset of ODVs with audio proposed in [9] and tested on an audio-visual saliency dataset proposed in [6] in mute, mono, and ambisonics modalities. Experimental results show that our model significantly outperforms other five state-of-the-art saliency models in three audio modalities. We discuss the results depending on the audio-visual contents and audio modalities of ODVs in detail.

The remainder of this paper is organized as follows.

978-1-7281-8068-7/20/$31.00

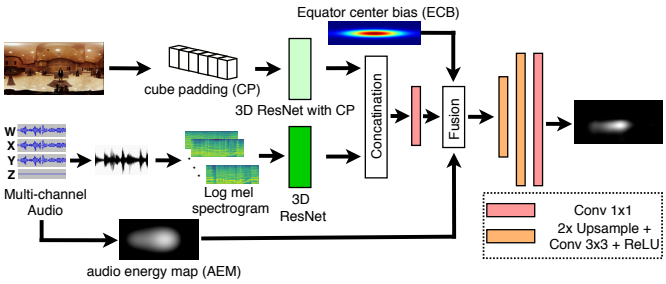

Fig. 1: Network architecture of the proposed AVS360 model.

Sec. II summarizes the related work on saliency prediction and Sec. III describes our proposed model. The experimental settings and results are presented in Sec. IV. Finally, Sec. V concludes the paper.

II. Related Work

Visual saliency has been broadly investigated in the literature [10]. Numerous saliency prediction algorithms have been proposed for standard 2D video and ODV. In this section, we briefly describe recent related works for ODV and audio-visual saliency for standard 2D video without the intention of providing comprehensive review on these topics.

Several algorithms have been proposed for modeling the visual attention of ODV. In particular, Salient360! Grand challenges at ICME 2017-2018 fostered saliency prediction models for omnidirectional image (ODI) and ODV by providing benchmark platforms and datasets [4]. In these challenges, Chao *et al.* [11], for instance, proposed *SalGAN360* using generative adversarial networks, and Monroy *et al.* [12] proposed *SalNet360* by expanding a traditional CNN-based saliency estimation algorithm for ODI. Furthermore, Cheng *et al.* [8] developed a spatial-temporal saliency prediction model trained in a weakly-supervised manner called *CubePading360*, and Xu *et al.* [9] proposed a large-scale eye-tracking ODV dataset with a gaze prediction model. Despite of the ODVs in their datasets are with audio, their proposed models ignored audio cues by considering only visual cues as an input.

Few recent works exist that incorporate both audio and visual cues in saliency modeling. For instance, Min *et al.* [13] proposed a multi-modal saliency model integrating the spatial, temporal, and sound features to predict saliency maps for standard 2D videos consist of high audio-visual correspondence scenes. Their results show that the audio signal contributes significantly to 2D video saliency prediction. A recent work by Tavakoli *et al.* [7] proposed a deep audio-visual embedding (DAVE) to investigate the applicability of audio cues in conjunction with visual ones in predicting saliency maps using deep neural networks. However, to the best of our knowledge, no research on modeling of audio-visual saliency prediction for ODV exists. To fill this gap in the literature, we propose a model that incorporates both omnidirectional visual and spatial audio cues to predict audio-visual saliency in ODV.

III. Proposed Model

The overall diagram of the proposed AVS360 model is illustrated in Fig. 1. In this work, we extend the previous

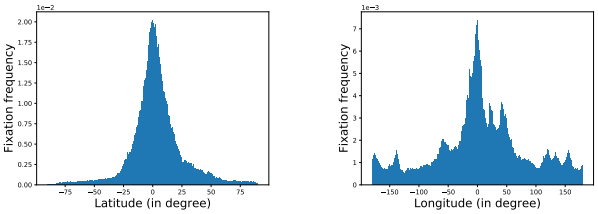

Fig. 2: Fixations distributions of our training set in latitude and longitude. Fixations scatter within the equator area along latitude and the center area in longitude in ODVs.

2D standard audio-visual saliency prediction model in [7] for ODV. The proposed model is an end-to-end 2-stream structure composed of two 3D ResNets to separately extract spatial-temporal visual and audio cues in ODV. A fusion network is appended after embedding of audio-visual features. The embedded features are then enhanced with the information of spatial audio (*i.e.*, ambisonics) and user navigation tendency (*i.e.*, ECB), and decoded into a saliency map. Each module is described in the following subsections.

A. Omnidirectional visual feature extraction

ODVs are captured as 3D spherical visual signals around a given center, and they are projected to 2D planar representations to be compatible with the existing video processing systems. ERP, which is a widely available format for ODV, contains geometric distortion along its latitude areas. To avoid this distortion, we adopt CP developed in [8] to alleviate geometric distortion in visual feature extraction. CP minimizes geometric distortion in ERP by using cubic projection to render 360° view on six cubic faces. These faces are concatenated with the connectivity of image padding between neighboring faces on the cube to avoid discontinuous image boundaries between cubic faces.

We employ CP into convolution and pooling layers in a 3D ResNet to extract spatial-temporal visual features of ODV. As visual saliency is not only guided by local visual stimuli but also overall 360° visual information, we also extract global spatial-temporal visual features in ERP format in the same 3D ResNet and combine them with local spatial-temporal visual features in CP format with an average pooling layer.

B. Spatial audio feature extraction

To extract audio features, we use a 3D ResNet in the other branch of our 2-stream structure. We follow the same audio processing procedures and parameters as used in DAVE [7]. Similarly, we convert audio signal resampled in 16KHz into successive overlaying frames of log mel-spectrograms. These frames are then used for extracting audio features with 3D ResNet.

C. Audio-visual feature embedding and fusion

After extracting audio and visual features via two 3D ResNets, we concatenate and embed audio and visual features with a 2D convolutional layer with 1×1 kernel size. We then fuse them with ECB and AEM to consider equator center viewing behavior and audio source location. Fig. 1 shows the heat map of ECB map and an example heat map of AEM.

We consider user viewing behaviour in ODV by generating an ECB with a 2D Gaussian distribution along longitude and latitude. In the light of the user navigation behavior analysis in the dataset [9], we observe that users' visual attention is concentrated within the equator area and the center area in ODVs. In this analysis, we see that viewers tend to look around horizontally more than vertically, and the visual stimuli often display a center bias in 360° conditions. These findings also agree with previous user studies [4], [5]. Fig. 2 shows the distributions of fixations for our training dataset [9].

We take spatial audio information of ODV into account by computing an AEM using audio directions in four channels (W, X, Y, Z) in ambisonics. AEM represents the audio energy distribution with a frame-by-frame heat map as in [2], [6]. We motivate the findings in [6], which indicates that perceiving sound sources directions could guide visual attention in ODVs.

We leverage an attention mechanism in residual form to incorporate the viewing tendency in ODVs [5] and the influence of sound source directions [6]. For this we first integrate ECB and AEM as $\hat{M} = (1 + AEM) \times ECB$, where $\hat{M}$ serves as attention map in the attention mechanism with residual connection: $\hat{F}^E = (1 + \hat{M}) \times F^E$, where F^E and $\hat{F}^E$ denote the audio-visual embedded features before and after modulated. With the residual connection, both the original features and the enhanced features are integrated and fed into the decoding blocks for generating saliency maps.

In decoding, we use two blocks which consist of a bilinear up-sampling layer with factor 2, and a 2D convolution layer with 3×3 kernel, and follow by batch normalization. A 2D convolution layer with 1×1 kernel is added in the end to generate a 1-channel gray-scale saliency map.

IV. Experiments

A. Experimental setup

1) Dataset: As we aim to extend audio-visual saliency model to ODVs with spatial audio, we used the dataset proposed by Chao *et al.* [6] as a test set. This dataset is the only existing audio-visual saliency dataset including visual saliency under mute, mono, and ambisonics modalities for ODV. It contains twelve ODVs where four are in the category Conversation, four are in the category Music, and four are in the category Environment. The category Conversation presents human talking, the category Music features people singing or instruments playing, while the category Environment contains background sound such as the noise of crowds or vehicles on the streets. We evaluate our model on these three categories under three audio modalities.

In training process, we used the dataset proposed by Xu *et al.* [9] as a training set. It is a large-scale eye-tracking dataset containing 208 ODVs with mono sound, while most of ODVs do not have correspondent audio-visual cues. Most of ODVs are with background narrator and music. Therefore, with the aim of modeling audio-visual saliency, we selected 27 ODVs with correspondent audio-visual cues equally allocated in the categories Conversation, Music, and Environment as our training set.

TABLE I: Mean values for saliency prediction accuracy of each component in our AVS360 model evaluated with the dataset [6]. (best in **bold** in each content category and audio modality)).

		mute		*mono*		*ambisonics*	
Cat.	**Models**	NSS	CC	NSS	CC	NSS	CC
Overall	*2stream o/w CP*	2.06	0.38	2.26	0.39	2.28	0.40
	2stream	2.22	0.41	2.44	0.42	2.46	0.43
	2stream+ECB	2.31	0.42	2.51	0.44	2.47	0.44
	2stream+ECB+AEM	**2.42**	**0.44**	**2.66**	**0.45**	**2.66**	**0.45**
Conver.	*2stream o/w CP*	2.24	0.40	2.56	0.41	2.37	0.38
	2stream	2.28	0.41	2.73	0.45	2.42	0.39
	2stream+ECB	2.41	0.44	2.82	0.45	2.40	0.40
	2stream+ECB+AEM	**2.57**	**0.47**	**3.12**	**0.50**	**2.68**	**0.42**
Music	*2stream o/w CP*	2.19	0.40	2.22	0.38	2.24	0.40
	2stream	2.30	0.42	2.37	0.39	2.42	0.46
	2stream+ECB	2.44	0.43	2.40	0.40	2.50	0.44
	2stream+ECB+AEM	**2.53**	**0.45**	**2.50**	**0.42**	**2.68**	**0.47**
Environ.	*2stream o/w CP*	1.74	0.34	2.00	0.37	2.21	0.41
	2stream	2.03	0.40	2.26	0.42	2.56	0.44
	2stream+ECB	2.07	0.39	2.31	0.42	2.59	0.46
	2stream+ECB+AEM	**2.16**	**0.41**	**2.37**	**0.43**	**2.62**	**0.47**

2) Implementation: We initialized our model with the weights of DAVE model pre-trained on 150 2D videos in mono sound providing diverse audio-visual contents with eye fixations. We followed the same structure of DAVE model and used 3D-ResNet-18 for extracting visual and audio features. In ECB, we follow the 68–95–99.7 rule in Gaussian distribution to design our ECB as shown in Fig. 2. We used a 2D Gaussian distribution map with $\sigma = 30°$ in latitude as about 95% fixations falls in $\pm 60°$ in latitude, and $\sigma = 75°$ in longitude as about 95% fixations falls in $\pm 150°$ in longitude. Due to the lack of the information of ambisonics in our training set, we only use AEM as auxiliary guidance in fusion procedure in test process. In training, Kullback-Leibler divergence (KLD) is used as a loss function. We set the batch size to six, initial learning rate to $1e{-}5$ with Adam optimizer. In total, we trained 20 epochs.

B. Contribution of each component in our proposed model

For evaluation, we selected two widely-used saliency evaluation metrics: normalized scanpath saliency (NSS), linear correlation coefficient (CC), to measure the performance of the prediction. The higher score of NSS and CC indicates a better prediction performance. Note that it is necessary to correct for the geometric distortions of oversampling areas closer to the north and south poles in ERP. Therefore, we follow [4] to weight down the oversampling areas with a sine function along latitude ($sin(y)$ for $y \in [0, \pi]$).

Table I lists the evaluation results of each component in our proposed model on the test dataset. We can see that the overall performances, including the performances in each content category and audio modality, are improved by using CP in visual feature extraction and embedding ECB and AEM in audio-visual features. In particular, the integration of spatial audio information (*i.e.*, AEM) not only improves the performances in ambisonics modality but also the performances in mute and mono modalities. Comparing three audio modalities, as our model uses sound information to assist the audio-visual saliency in ODVs, the overall performances of mono modality and ambisonics modality are better than that in mute modality.

TABLE II: Mean values for saliency prediction accuracy of the state-of-the-art models evaluated with the dataset [6] (best in **bold** in each audio modality and content category).

		mute		*mono*		*ambisonics*	
Cat.	**Models**	NSS	CC	NSS	CC	NSS	CC
Overall	*SalNet360 [12]*	1.49	0.29	1.55	0.28	1.47	0.26
	SalGAN360 [11]	1.58	0.31	1.65	0.30	1.60	0.30
	CP360 [8]	1.16	0.24	1.19	0.23	1.16	0.22
	MMS [13]	1.24	0.25	1.39	0.25	1.35	0.25
	DAVE [7]	1.92	0.36	2.16	0.38	2.13	0.38
	AVS360 (Ours)	**2.42**	**0.44**	**2.66**	**0.45**	**2.66**	**0.45**
Conver.	*SalNet360 [12]*	1.72	0.33	1.84	0.31	1.61	0.28
	SalGAN360 [11]	1.86	0.36	1.94	0.33	1.77	0.31
	CP360 [8]	1.20	0.24	1.25	0.22	1.19	0.22
	MMS [13]	1.53	0.30	1.91	0.33	1.70	0.30
	DAVE [7]	2.18	0.40	2.68	0.44	2.25	0.37
	AVS360 (Ours)	**2.57**	**0.47**	**3.12**	**0.50**	**2.68**	**0.42**
Music	*SalNet360 [12]*	1.46	0.27	1.48	0.28	1.40	0.25
	SalGAN360 [11]	1.55	0.29	1.52	0.29	1.53	0.28
	CP360 [8]	1.15	0.23	1.14	0.22	1.14	0.22
	MMS [13]	0.99	0.19	0.96	0.17	1.03	0.20
	DAVE [7]	1.67	0.32	1.66	0.30	1.93	0.36
	AVS360 (Ours)	**2.53**	**0.45**	**2.50**	**0.42**	**2.68**	**0.47**
Environ.	*SalNet360 [12]*	1.30	0.28	1.33	0.27	1.39	0.27
	SalGAN360 [11]	1.33	0.29	1.47	0.29	1.51	0.30
	CP360 [8]	1.12	0.24	1.17	0.23	1.18	0.23
	MMS [13]	1.18	0.24	1.30	0.26	1.30	0.25
	DAVE [7]	1.89	0.36	2.16	0.39	2.21	0.41
	AVS360 (Ours)	**2.16**	**0.41**	**2.37**	**0.43**	**2.62**	**0.47**

We also notice that there is no significant difference between mono and ambisonics modality in their overall performances.

We then take a more in-depth look at the evaluation of each content category. In the category Conversation, the mono modality has better accuracy performance than ambisonic modality even when ambisonics information (*i.e.*, AEM) is provided. In contrast, in category Music and Environment, ambisonics modality outperforms mono modality in the model with and without AEM. These results are inconsistent with the finding of the user behavior analysis mentioned in [6], where users may tend to follow the human voice in categories Conversation and Music while looking around in general in category Environment. The reason might be the inefficiency of the integration method (*i.e.*, attention mechanism in the residual form) of AEM and ECB maps. The lack of a large-scale audio-visual saliency dataset of ODVs with spatial sound disables us for developing a trainable integration method. We will leave this as our future work.

C. Comparison to the state of the arts

Table II shows the accuracy of our proposed AVS360 model and other five state-of-the-art models evaluated on the test dataset. SalGAN360 [11] and SalNet360 [12] are visual saliency prediction for ODIs, CP360 [8] is visual saliency prediction for ODVs, while MMS [13] and DAVE [7] are audio-visual saliency prediction for 2D videos. We can see from the results that our model significantly outperforms the others in overall ODVs and every content category in all three audio modalities. To our surprise, even in mute modality, our model, which takes account of spatial auditory information, surpasses a significant margin over the models, only taking account of visual information. The results show that using AEM as an input feature can improve the saliency prediction models for ODV.

V. Conclusion

In this paper, we proposed an end-to-end trainable audio-visual saliency prediction model for ODVs with spatial audio. The experimental results show that our proposed model significantly outperforms other five state-of-the-art methods in mute, mono, and ambisonics modalities. The results also suggest that using audio energy map computed from ambisonics as an input feature could benefit the saliency prediction models for ODV. In addition, we demonstrated that the 360° representation with cube padding, equator center bias and audio energy map contribute significantly to audio-visual saliency prediction in ODVs. As a future work, we plan to improve the fusion block of the proposed model and propose a large scale audio-visual ODV dataset with spatial audio.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.

References

[1] A. Rana, C. Ozcinar, and A. Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[2] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," in *Advances in Neural Information Processing Systems*, 2018.

[3] C. Ozcinar, J. Cabrera, and A. Smolic, "Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, March 2019.

[4] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, ser. MMSys '18, 2018.

[5] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018.

[6] F. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2020.

[7] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, "DAVE: A deep audio-visual embedding for dynamic saliency prediction," *CoRR*, vol. abs/1905.10693, 2019. [Online]. Available: http://arxiv.org/abs/1905.10693

[8] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360° immersive videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[10] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[11] F. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2018.

[12] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "Salnet360: Saliency maps for omni-directional images with cnn," *Signal Processing: Image Communication*, vol. 69, 2018, salient360: Visual attention modeling for 360° Images.

[13] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.