ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

Video Saliency Detection Based on Audio Visual Consistency

Meng Yuan[1, a, *], Yuxiao Yu[2, b]

[1]College of computer science and technology, Qingdao University, Qingdao, 266000, China

[2]College of computer science and technology, Qingdao University, Qingdao, 266000, China

[a]yuanmeng_926@163.com, [b]yuxiaoyu_yu@163.com

Abstract

Saliency detection is an important research content of computer vision, especially video saliency detection has received more attention. It aims to automatically identify the object or area that attracts human attention in the scene by simulating human vision system, which can help people obtain important information from massive data and allocate limited computing resources to more important information. However, we humans perceive sound and visual modes at the same time, so the Saliency detection of audio visual combination is more in line with the real scene. However, most of the existing audio-visual fusion detection algorithms use dual stream structure to extract the features of audio and video information respectively, and then simply fuse the audio-visual features to obtain the final prediction map. Then the audio information and visual information in the data set will be irrelevant. Therefore, when the audio and visual features are inconsistent, the direct fusion of audio and visual features will have a negative impact on the visual features. Therefore, this paper proposes video saliency detection based on audio-visual consistency.

Keywords

Saliency detection; Multimodal fusion; Selective fusion.

1. INTRODUCTION

The purpose of saliency detection [1] is to let the computer simulate human visual characteristics through intelligent algorithms and extract the significant regions in the image (i.e. the regions of human interest). Visual attention is a selection process and an intelligent mechanism of human visual system. It allows you to select the most attractive area in the visual scene. Saliency detection is divided into saliency object detection [2] and saliency fixation detection [3], as shown in Figure 1. In this paper, we focus on the latter research. The saliency detection mentioned in this paper is significance concern detection. At present, a lot of work has been done on the research of image saliency, but less attention has been paid to the research of video saliency detection, and the research of audio and video saliency is less concerned. Although there is evidence [4] that there is a strong correlation between auditory and visual cues and their common contribution to attention, so far, most video saliency models ignore audio cues and rely heavily on spatio-temporal visual cues as the information source of saliency detection.

In life scenes, when the audio and visual stimuli are consistent, it will attract our attention more. For example, there are two entities in our sight, dog and cat, and our attention distribution is almost equal, but when there is the audio stimulation of dog barking, we will pay more attention to the dog. Therefore, the consistency between learning audio and video plays a certain role in promoting the computer to better simulate the characteristics of human

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

attention distribution in deep learning. Arandjelovic [5] first defined the audio-visual consistency learning task (AVC) in their work, AVC task is a simple binary classification task: given a video frame example and an audio, judge whether they correspond. Matched (positive) audio-visual pairs are extracted from the same video at the same time, while unmatched (negative) pairs are extracted from different videos. Its work is realized by learning and detecting various semantic concepts of video and audio segments in an unsupervised way, and the semantic information that can be learned is found after visualizing the features learned by the middle layer of video network branches.

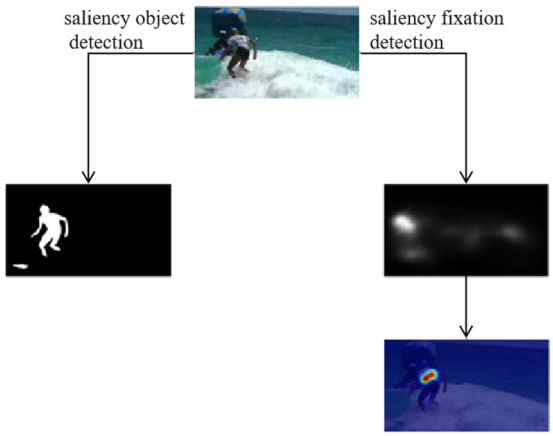

Figure 1. Salient object detection and salient fixation detection

Senocak et al. [6] proposed a dual stream network structure with attention mechanism, which can find objects consistent with sound on the picture according to the input of different audio and picture pairs. For example, for a picture containing human and car scenes, when the input audio is human voice, the attention mechanism will pay attention to the human area; The input audio is the sound of a motor vehicle. The attention mechanism will make the part of the car in the picture allocate more attention. The attention model in its work will produce different weights in different areas according to the interaction between audio and video, and focus on the areas consistent with audio in the picture according to the score of the weight.

2. METHOD

2.1. Network Structure

Different from the current algorithm that only considers the fixation prediction of visual modal design, we consider the impact of audio on video significance detection. We use the dual stream method to calculate the audio and visual features respectively. As shown in Figure 2, the overall network framework diagram is divided into four parts: spatio-temporal visual feature extraction, audio feature extraction, audio-visual fusion mechanism and saliency feature calculation.

In this paper, 3D ResNet framework is used to calculate spatiotemporal visual features, and 3D convolution can effectively encode spatiotemporal information. In addition, it is light-weight and pretrained on large datasets, which makes it able to carry out migration learning quickly and effectively. The video network branch includes four ResNet convolution blocks conv_ 1, conv_ 2, conv_ 3, conv_ 4. Output F_1, F_2, F_3 and F_4 on different temporal and spatial scales, such as formula (1), cascade the features of receptive fields of different sizes to obtain the final visual features.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

$$F_m = conv3D(input, conv\_m), m = 1,2,3,4 \quad (1)$$

The audio is cropped to match the duration of visual frames (i.e. 16 frames). The network adopted by the audio branch is the network model of the first seven layers of SoundNet. The advantage of the convolution kernel designed in the network is that it can process audio information of variable length, so it does not need sub sampling to deal with the change of sampling rate between different videos. For the preprocessing of original audio data, a Hamming window is applied to give higher weight to the central audio value at the current time, and give corresponding participation weight to the past and future audio frequency values. After inputting the preprocessed audio data into network for feature extraction, these layers are followed by a maximum time pooling layer, which is used to obtain the audio feature vector with fixed dimension of the whole sequence, as shown in Formula (2).

$$F_a = conv(input, w_a) \quad (2)$$

Where, conv() represents one-dimensional convolution from audio data to corresponding advanced audio features. w_a represents the parameters in the process of obtaining audio features.

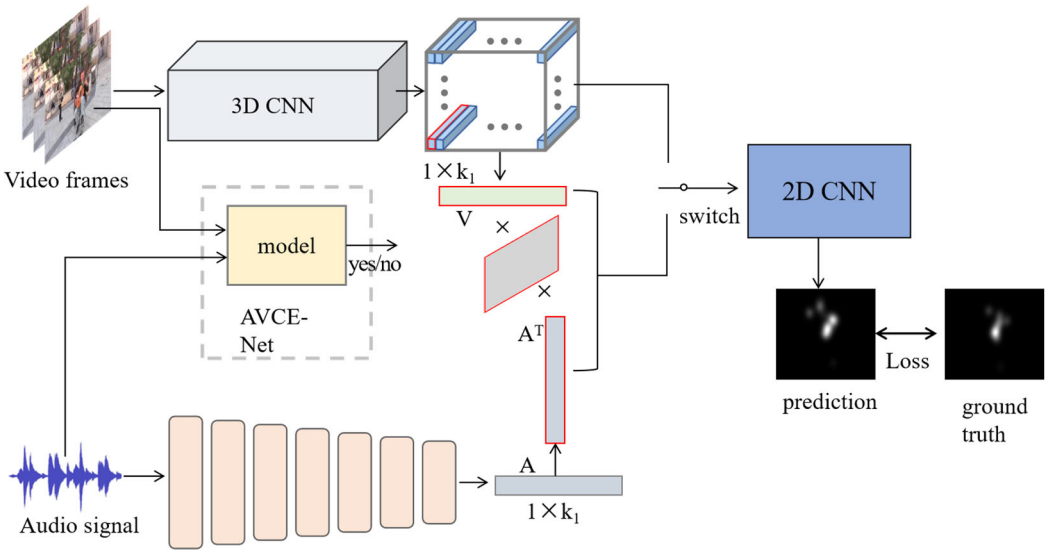

Figure 2. Network structure

2.2. Multi Modal Fusion

In this paper, a bilinear interpolation transformation fusion algorithm is proposed to fuse audio-visual features. The calculation process is shown in Formula (3). The result of audio-visual fusion is expressed in S_{av}, M represents two-dimensional matrix, F_v is visual features and b is random parameters. The bilinear interpolation transformation fusion scheme proposed in this paper has a significant advantage: it does not need the visual and audio saliency to have the same dimension size, and the dimension transformation can deal with the size mismatch by the transformation matrix M. However, this fusion also has its own limitations, that is, the semantic correspondence between audio-visual channels will be destroyed, which makes it very difficult to model the complex audio-visual information interaction. Therefore, our base network selects ResNet model, which can provide semantic information for the extracted features. The learned semantic information can effectively reduce the problem domain. Therefore, even in the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

complex video and audio environment, it can easily achieve the fusion state of audio and video complementarity.

$$S_{av} = F_a^T M F_v \tag{3}$$

2.3. Saliency Calculation

After the audio-visual fusion results are obtained by bilinear interpolation transformation algorithm, the fusion features are further processed, and the bilinear up sampling block with input factor of 2 is used. The purpose of up sampling is to enlarge the fusion features, so as to match the size of the predicted result image with the original size. Finally, the convolution layer is used to output the final significance prediction map.We choose the most widely used Kullback Leibler (KL) divergence as the loss function, as shown in Formula (4). It is often used in significance prediction tasks to calculate the difference between two probability distributions, including fixation prediction graph(S) and ground truth graph(G), which φ represent the regular term. When the value is smaller, the difference between and is smaller, which makes it closer through multiple training.

$$KL(S,G) = \sum_i G_i \log\left(\frac{G_i}{S_i + \varphi} + \varphi\right) \tag{4}$$

3. EXPERIMENTAL RESULTS

3.1. Datasets

Deep learning is inseparable from the support of big data. Data set is very important in the comparison of algorithms and the performance verification of results. At present, there are six public data sets in the field of audiovisual significance detection, including DIEM [7], AVAD [8], Coutrot1 [9], Coutrot2 [10], SumMe [11] and ETMD [12]. Different from the traditional visual saliency detection set, the fixation point collection of the audio-visual saliency detection data set is carried out in the audio-visual environment, while the traditional fixation point collection is carried out only under the visual conditions, as shown in Table 1. The number of segments and frames of each data is shown in detail. From the table, it can be seen that the total number of frames of the six data sets is very large.

Table 1. Audio-visual datasets information

Datasets	Date	Clips	Frames
DIEM	2010	84	78167
AVAD	2016	45	9564
Coutrot1	2013	60	25223
Coutrot2	2014	15	17134
SumMe	2019	25	109788
ETMD	2019	12	52744

3.2. Metrics

In order to evaluate the robustness of the model fairly, it is necessary to synthesize a variety of evaluation metrics for quantitative analysis. The five indexes are linear correlation coefficient [13] (CC), similarity measure [14] (SIM), normalized scanpath saliency [15] (NSS), sAUC [16] and AUC_ J [17] .

(1) Linear correlation coefficient

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

Linear correlation coefficient is a method to measure the linear correlation between the results of the saliency map of human eye focus predicted by the model and the true value. The calculation method is shown in formula (5):

$$CC(S,G)=\frac{\mathrm{cov}(S,G)}{\sigma(S)\times\sigma(G)} \tag{5}$$

(2) Similarity

It is to calculate the similarity between the predicted saliency map result and the ground truth, normalize the predicted result and the true value, calculate the minimum value on each pixel, and finally add it, as shown in formula (6):

$$SIM=\sum_i\left(S(i),G(i)\right) \tag{6}$$

(3) Normalized scanpath saliency

NSS is an metric dedicated to the detection of the human eyefixation, N is the total number of human eye positionan, as shown in formula (7):

$$NSS=\frac{1}{N}\sum_i^1 S(i)*G(i) \tag{7}$$

3.3. Comparison Results with Mainstream Methods

In this section, We quantitatively compare the audio-visual significance prediction method proposed in this paper with the representative methods of video significance prediction considering only visual features SBF [18], WSS [19], MWS [20], WSSA [21], ACLNet [22], Deep VS[23] and audio-visual significance detection model DAVE [24] on six public data sets, as shown in Table 2.It can be seen from the table that our video saliency detection performance shows the best advantage on the whole.

At the same time, we compare our method with the existing audio-visual significance detection method DAVE. Because the processing method of audio branch in DAVE method is the same as that of video branch, the convolutional neural network can not extract audio features efficiently. Therefore, this paper uses different processing methods for different modal information to show its advantages. From the table, we can also see that the value of Deep VS method in metrics CC, SIM and NSS is more prominent on the AVAD datasets. The reason is closely related to the content of the data set. The objects in the video scene of this data set are relatively single, and the background in most video segments is not messy. Even the video significance detection method without considering audio information will achieve high detection effect. On other data sets with multiple objects and chaotic background, the method of considering audio information will help to quickly focus on the vocal object in the case of multiple objects.

In addition, we also calculated the single V (OUR-V) prediction result of our network, that is, the prediction result without considering the significance of audio information. As shown in Table 3, it can be seen from the table that our video saliency detection performance shows the best advantages on the whole. Compared with the OUR-V result of our method, adding audio information improves the detection effect. However, the result of OUR-V is better than OUR-AV in some metrics. The reason may be that the accuracy of audio-visual consistency judgment is not so high.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

Table 2. The comparison results with mainstream method

DataSet / Methods	DIEM					SumMe				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
SBF	0.296	1.092	0.760	0.610	0.302	0.195	0.864	0.761	0.573	0.203
WSS	0.348	1.301	0.806	0.623	0.339	0.273	1.272	0.821	0.598	0.254
MW	0.355	1.325	0.808	0.631	0.336	0.249	1.125	0.813	0.596	0.228
WSSA	0.324	1.223	0.776	0.615	0.310	0.197	0.924	0.740	0.572	0.198
DeepVS	0.452	1.860	0.841	0.626	0.392	0.317	1.620	0.842	0.612	0.262
ACLNet	0.523	2.020	0.869	0.622	0.428	0.380	1.790	0.869	0.609	0.297
DAVE	0.544	2.165	0.873	0.635	0.474	0.361	1.640	0.866	0.592	0.324
OUR-AV	0.594	2.295	0.893	0.735	0.496	0.432	1.998	0.890	0.651	0.353
DataSet / Methods	Coutrot1					Coutrot2				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
SBF	0.176	0.696	0.716	0.525	0.218	0.174	1.416	0.857	0.603	0.168
WSS	0.214	0.836	0.766	0.538	0.244	0.190	1.248	0.844	0.590	0.194
MW	0.186	0.713	0.732	0.532	0.232	0.172	1.297	0.855	0.614	0.169
WSSA	0.166	0.728	0.698	0.536	0.185	0.213	1.614	0.858	0.599	0.192
DeepVS	0.360	1.770	0.831	0.562	0.317	0.449	3.790	0.926	0.647	0.259
ACLNet	0.425	1.920	0.850	0.543	0.361	0.449	3.160	0.927	0.594	0.323
DAVE	0.440	2.011	0.856	0.563	0.390	0.643	4.860	0.956	0.682	0.462
OUR-AV	0.426	1.865	0.867	0.593	0.387	0.712	5.431	0.957	0.715	0.464
DataSet / Methods	ETMD					AVAD				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
SBF	0.231	1.124	0.776	0.620	0.212	0.260	1.181	0.800	0.557	0.257
WSS	0.313	1.523	0.842	0.644	0.261	0.312	1.363	0.822	0.565	0.282
MW	0.289	1.403	0.831	0.647	0.235	0.268	1.168	0.806	0.559	0.262
WSSA	0.220	1.069	0.792	0.622	0.200	0.268	1.233	0.781	0.570	0.257
DeepVS	0.462	2.480	0.904	0.686	0.350	0.581	3.170	0.905	0.560	0.446
ACLNet	0.477	2.360	0.915	0.675	0.329	0.528	3.010	0.897	0.586	0.391
DAVE	0.456	2.301	0.912	0.655	0.380	0.486	2.479	0.893	0.553	0.407
OUR-AV	0.581	2.881	0.934	0.735	0.438	0.489	2.595	0.913	0.609	0.421

Table 3. Ablation experiment

DataSet / Methods	DIEM					SumMe				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
OUR-V	0.529	2.041	0.869	0.644	0.454	0.417	1.885	0.895	0.631	0.344
OUR-AV	0.594	2.295	0.893	0.735	0.496	0.432	1.998	0.890	0.651	0.353
DataSet / Methods	Coutrot1					Coutrot2				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
OUR-V	0.386	1.634	0.841	0.557	0.353	0.626	4.179	0.939	0.681	0.470
OUR-AV	0.426	1.865	0.867	0.593	0.387	0.712	5.431	0.957	0.715	0.464
DataSet / Methods	ETMD					AVAD				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
OUR-V	0.552	2.767	0.936	0.709	0.409	0.458	2.557	0.902	0.560	0.431
OUR-AV	0.581	2.881	0.934	0.735	0.438	0.489	2.595	0.913	0.609	0.421

4. CONCLUSION

Aiming at the problem that the current audio-visual human eye fixation detection fails to consider the audio-visual consistency, this paper proposes a video saliency model based on audio-visual consistency. The model uses the form of two streams to extract the audio-visual

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

features respectively, and finally obtains the prediction results after fusion. The experimental results show that considering the significance of audio information, the detection effect is better. The future research direction is to improve the accuracy of audio-visual consistency judgment, retrain the network, and further improve the performance of audio-visual human eye focus detection.

REFERENCES

[1] Borji A. Saliency prediction in the deep learning era: An empirical investigation[J]. arXiv preprint arXiv:1810.03716, 2018, 10.

[2] Bak C, Kocak A, Erdem E, et al. Spatio-temporal saliency networks for dynamic saliency prediction[J]. IEEE Transactions on Multimedia, 2017, 20(7): 1688-1698.

[3] Alshawi T, Long Z, AlRegib G. Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection[J]. IEEE Transactions on Image Processing, 2018, 27(6): 2818-2827.

[4] Chen Y, Nguyen T V, Kankanhalli M, et al. Audio matters in visual attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(11): 1992-2003.

[5] Arandjelovic R, Zisserman A. Look, listen and learn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 609-617.

[6] Senocak A, Oh T H, Kim J, et al. Learning to localize sound source in visual scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4358-4366.

[7] Mital P K, Smith T J, Hill R L, et al. Clustering of gaze during dynamic scene viewing is predicted by motion[J]. Cognitive computation, 2011, 3(1): 5-24.

[8] Min X, Zhai G, Gu K, et al. Fixation prediction through multimodal analysis[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2016, 13(1): 1-23.

[9] Coutrot A, Guyader N. How saliency, faces, and sound influence gaze in dynamic social scenes[J]. Journal of vision, 2014, 14(8): 5-5.

[10] Coutrot A, Guyader N. Multimodal saliency models for videos[M]//From Human Attention to Computational Attention. Springer, New York, NY, 2016: 291-304.

[11] Gygli M, Grabner H, Riemenschneider H, et al. Creating summaries from user videos[C]//European conference on computer vision. Springer, Cham, 2014: 505-520.

[12] Koutras P, Maragos P. A perceptually based spatio-temporal computational framework for visual saliency estimation[J]. Signal Processing: Image Communication, 2015, 38: 15-31.

[13] Borji A, Itti L. State-of-the-art in visual attention modeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 185-2073.

[14] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[15] JBorji A, Sihite D N, Itti L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study[J]. IEEE Transactions on Image Processing, 2012, 22(1): 55-69.

[16] Bylinskii Z, Judd T, Oliva A, et al. What do different evaluation metrics tell us about saliency models?[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(3): 740-757.

[17] Judd T, Durand F, Torralba A. A benchmark of computational models of saliency to predict human fixations[J]. 2012.

[18] Zhang D, Han J, Zhang Y. Supervision by fusion: Towards unsupervised learning of deep salient object detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4048-4056.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202205_8(5).0024

[19] Wang L, Lu H, Wang Y, et al. Learning to detect salient objects with image-level supervision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[20] Zeng Y, Zhuge Y, Lu H, et al. Multi-source weak supervision for saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6074-6083.

[21] Zhang J, Yu X, Li A, et al. Weakly-supervised salient object detection via scribble annotations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12546-12555.

[22] Wang W, Shen J, Guo F, et al. Revisiting video saliency: A large-scale benchmark and a new model[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4894-4903.

[23] Jiang L, Xu M, Liu T, et al. Deepvs: A deep learning based video saliency prediction approach[C]//Proceedings of the european conference on computer vision (eccv). 2018: 602-617.

[24] Tavakoli H R, Borji A, Rahtu E, et al. Dave: A deep audio-visual embedding for dynamic saliency prediction[J]. arXiv preprint arXiv: 1905.10693, 2019.