SST-Sal: A Spherical Spatio-Temporal Approach for Saliency Prediction in 360º Videos

Edurne Bernal, Daniel Martín, Diego Gutiérrez, Belén Masiá

Afilliation: Graphics and Imaging Lab (GILab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: edurnebernal@unizar.es

Abstract

We present a deep learning approach to visual attention prediction in 360º videos. We resort to recurrent neural networks to model the inherent spatio-temporal features of visual behavior, while tailoring our model to the particularities of 360º content. Our model outperforms previous state-of-the-art works by a large margin.

Introduction

In the last years, virtual reality (VR) is gaining traction as a powerful tool in many fields, such as the manufacturing industry, online marketing, architecture, design, or education. However, much remains unknown about the grammar and visual language of this new medium: Understanding and predicting how humans behave in virtual environments, while crucial for all those applications, remains an open problem.

Previous works have attempted to analyze human visual behavior through recorded gaze data from real observers. While there is significant variability among individuals, the data also reveals the existence of some common patterns among observers [1]. This common behavior is strongly influenced by the areas of interest that capture human visual attention (i.e., the salient regions of the scene). Therefore, a first step in understanding human visual behavior in virtual environments relies on modeling how the different regions in a scene attract users' attention. This is often represented with a saliency map of the scene.

Saliency prediction for traditional 2D content has already been widely explored in the last decades. However, the patterns and visual behaviors known for 2D do not necessarily hold for VR content [2].

In this work, we present SST-Sal, a novel saliency prediction model for 360º videos, based on a deep neural network. A key aspect of our model is that, in contrast to previous works, it accounts for *both* temporal and spatial information at feature extraction time, with an encoder and decoder built over a recurrent neural network. To further favor temporal feature learning, we provide the network with optical flow estimations of consecutive frames. Finally, we present a novel spherical Kullback-Leibler Divergence loss function specifically tailored to 360º content. While here we target saliency prediction, our design choices could also benefit other applications dealing with 360º videos. Our evaluations show that SST-Sal outperforms previous state-of-the-art works, yielding results that resemble human visual behavior more closely. Our work has been accepted to CEIG 2022 and referred to the journal Computers & Graphics [3].

Our Model

SST-Sal is based on a type of recurrent neural networks termed Long Short-Term Memory (LSTMs) cells, which are able to infer both the spatial and temporal relationships between video frames. We provide those LSTM cells with spherical convolutions to handle the distortion introduced by the equirectangular representation of 360º video frames. Our model architecture follows an encoder-decoder approach, where both elements are based on the aforementioned LSTMs. Besides the 360º RGB frames, we feed our model with optical flow estimations between consecutive frames to learn the relationships between motion and saliency.

We train our model with a modified version of the traditional Kullback-Leibler Divergence (KLDiv) loss function. Our spherical KLDiv accounts for the distortion introduced by the sphere-to-plane projection by weighting the contribution of each pixel to the loss by its solid angle. SST-Sal was trained in a selection of videos from the VR-EyeTracking dataset [4].

Results and Evaluation

Some qualitative results can be seen in Figure 1. The predicted saliency maps yielded by our model resemble the ground truth, focusing on small, yet relevant regions of the scene. Additionally, we compare our results to those of previous state-of-the-art approaches with a set of metrics traditionally used in saliency assessment, in two test sets. Our model outperforms them by a large margin (see Table 1).

We conducted five ablation studies, analyzing: the influence of the input data resolution, the use of spherical convolutions, the suitability of the loss function, the inclusion of optical flow, and the advantages reported by the Spherical ConvLSTMs. These studies endorse the decisions adopted in the design stage of our model.

Tabla 1. Quantitative comparisons of the proposed model against ATSal [5], CP-360 [6], and Martin et al.'s [7], for two different datasets. Arrows indicate whether higher or lower is better. The values represent the mean score among the different videos in the dataset for three different metrics, and in brackets is shown the averaged standard deviation.

	VR-EyeTracking dataset			Sports-360 dataset		
	CC ↑	SIM ↑	KLDiv ↓	CC ↑	SIM ↑	KLDiv ↓
ATSal*	0.298	0.216	9.858	0.246	0.184	10.552
	(0.087)	(0.041)	(1.000)	(0.090)	(0.050)	(1.221)
CP-360	0.229	0.148	10.665	0.228	0.135	11.753
	(0.049)	(0.025)	(0.556)	(0.0.055)	(0.031)	(0.731)
Martin en al.'s	0.138	0.152	8.610	0.240	0.183	11.191
	(0.054)	(0.033)	(0.941)	(0.078)	(0.049)	(1.176)
Our Model	**0.500**	**0.338**	**7.371**	**0.439**	**0.284**	**8.610**
	(0.123)	**(0.058)**	**(1.218)**	**(0.143)**	**(0.070)**	**(1.591)**

REFERENCES

[1]. MARTIN, D., SERRANO, A., BERGMAN, A., WETZSTEIN, G., and MASIA, B. ScanGAN360: A Generative Model of Realistic Scanpaths for 360 Images. *IEEE Transactions on Visualization and Computer Graphics.* 2022.

[2]. SITZMANN, V., SERRANO, A., PAVEL, A., AGRAWALA, M., GUTIERREZ, D., MASIA, B., and WETZSTEIN, G. How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics.* 2017.

[3]. BERNAL, E., MARTIN, D., GUTIERREZ, D., and MASIA, B. SST-Sal: A Spherical Spatio-Temporal Approach for Saliency Prediction in 360◦ Videos. *Spanish Computer Graphics Conference.2022.*

[4]. XU, Y., DONG, Y., WU, J., SUN, Z., SHI, Z., YU, J., GAO, S.: Gaze prediction in dynamic 360° immersive videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2018.

[5]. DAHOU, Y., TLIBA, M., MCGUINNESS, K., and O'CONNOR, N. ATSal: An Attention Based Architecture for Saliency Prediction in 360 Videos *Lecture Notes in Computer Science.* 2020.

[6]. CHENG, H., CHAO, C., DONG, J., WEN, H.K., LIU, T.L., and SUN, M. Cube Padding for Weakly-Supervised Saliency Prediction in 360º Videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2018.

[7]. MARTIN, D., SERRANO, A., MASIA, B. Panoramic convolutions for 360º single-image saliency prediction. *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*. 2020.

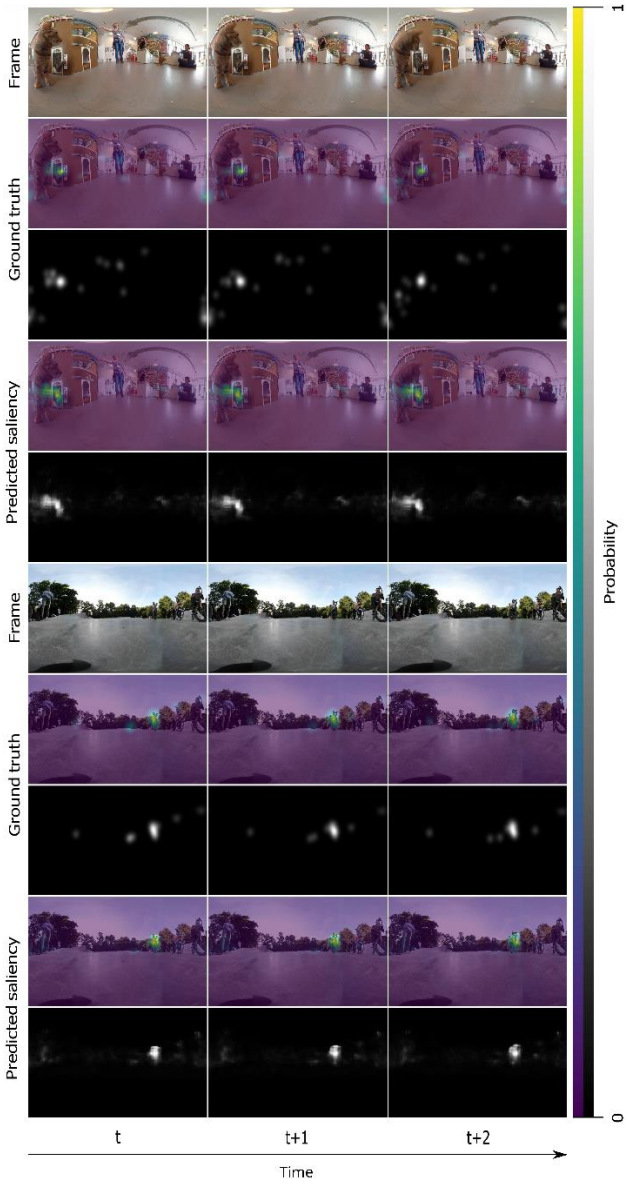

Figure 2. Results obtained with our model for two video sequences. The horizontal axis represents time. Rows show the input frames, the ground truth saliency, and the predicted saliency of the sequences. Saliency is represented both in greyscale and as a heat map blended with the frame's image, where warmer colors indicate more salient areas.