SalyPath360: Saliency and Scanpath Prediction Framework for Omnidirectional Images

Mohamed Amine KERKOURI¹, Marouane TLIBA¹, Aladine CHETOUANI¹, Mohamed SAYEH²
¹Laboratoire PRISME, Université d’Orléans, Orléans, FRANCE
²University of Oran 1 ORAN, ALGERIA

Abstract
This paper introduces a new framework to predict visual attention of omnidirectional images. The key setup of our architecture is the simultaneous prediction of the saliency map and a corresponding scanpath for a given stimulus. The framework implements a fully encoder-decoder convolutional neural network augmented by an attention module to generate representative saliency maps. In addition, an auxiliary network is employed to generate probable viewport center fixation points through the SoftArgMax function. The latter allows to derive fixation points from feature maps. To take advantage of the scanpath prediction, an adaptive joint probability distribution model is then applied to construct the final unbiased saliency map by leveraging the encoder decoder-based saliency map and the scanpath-based saliency heatmap. The proposed framework was evaluated in terms of saliency and scanpath prediction, and the results were compared to state-of-the-art methods on Salient360! dataset. The results showed the relevance of our framework and the benefits of such architecture for further omnidirectional visual attention prediction tasks.

1. Introduction
Virtual Reality (VR) applications provide high quality of immersive user experiences. Most of VR applications are in the form of 360 video, whereas the frames are represented under a new format of multimedia content called omnidirectional image. These images cover the whole spherical viewing space (360° × 180°), where the user has the freedom of attending to any direction just by pointing his head to any direction. The viewport of the 360° image is defined by the device-specific viewing angle (typically 120 degrees), which delimits horizontally the scene from the head direction center, called viewport center. The rendering of 360° viewport of the images is supported by many types of sphere to plane coordinates mapping transformations, EquiRectangular Projection (ERP) is one of the most widely used formats of uniform quality mapping projection [15][14]. It projects the spherical content to a single high resolution 2D plane, where the longitudinal and latitudinal spherical coordinates are represented on the horizontal and vertical ERP axes, respectively.

Unlike the traditional fixed viewport delivery of 2D content, the immersive experience is delivered using recent technologies such as Head Mounted Display (HMD). They are empowered by the ability to investigate spherical space, enabling them to have the best realistic immersive experience with a high consumption of resources. Therefore, the capacity to predict the attended viewport that corresponds to the orientations of the head movements beforehand, helps to optimize the delivery process and to provide a higher Quality of Experience (QoE) to the viewer [3].

This can be achieved through the prediction of the human visual attention that reflects the most interesting regions within the field of view of the users. This natural mechanism allows humans to explore complex scenes effortlessly and devotes their limited perceptual resources to the most pertinent subsets of received sensory information [24]. The attractive regions, often called salient regions, are usually represented in a heatmap (i.e saliency map). This map models the distribution of the gaze fixations describing the probability that a pixel is salient. The saliency maps are generated by processing the scanpaths of different viewers, which are defined as a sequence of successive fixation points of the viewer’s gaze while exploring the image [19].

Unlike conventional 2D images, the users are exposed in omnidirectional images to a larger degree of freedom. Visual attention modeling in 360° content is conducted by predicting the center of most probable attended viewports, which reflect the trajectory of the head movements. Moreover, we assume as in [25] that human attention toward omnidirectional content is governed by some statistical bias, as they tend more to equatorial and frontal regions than others, referred as equator bias.

Studies on omnidirectional content were pioneered by the work of Bogdanova et al. [4, 5] where the spherical static saliency map was generated by normalizing and merging chromatic, intensity and three cue conspicuities. They also created a motion spherical saliency map through a motion pyramid decomposition. At the last stage, the two resulting maps were fused to produce the spherical frame saliency map. In [12], the authors introduced a Fused Saliency Map (FSM) to predict visual attention on ERP omnidirectional images by adopting the well-known 2D saliency model SALICON [18]. In [20], the authors extended both the 2D Boolean Map Saliency (BMS) [27] and GBVS [17] by incorporating the characteristics of ERP images.

With the high performance of deep neural networks in imaging, various saliency prediction models were introduced. In [22], the authors fine-tuned a 2-Dimensional static saliency model, called PanoSalNet, on ERP frames for the task of head movement prediction. The resulting saliency map is further enhanced by prior statistical bias. In [5], the authors fine-tuned also a 2D static model, called SalGAN, on 360° image dataset using cube-map projection and a new objective function, leveraging the combination of three saliency measures. Instead of using a supervised approach to learn saliency from labeled data, In [13], the authors proposed a novel attention-based architecture that adapts the encoded latent vector to the characteristics of omnidirectional images through the extended receptive field. They exploited a Cubic Map Projection (CMP) to improve prediction on polar regions. In
In [11], the authors proposed a new framework that applies existing 2D saliency models on ERP images without requiring in-depth adaptation of the prediction algorithms. They adopt an adaptive weighted joint probability distribution on different kinds of projection of omnidirectional images.

Scanpath prediction methods are more scarce in literature, even more so in 360° methods. In [29], the authors proposed a model that uses low level features to produce a saliency map. The resulting map is then binarized and fixation points are generated from the obtained binary map using a clustering method. A graph is constructed from the fixation points. The scanpath generated by maximizing the sum of transfer probabilities on the graph edges. In [2], the authors proposed a hybrid method, that uses deep neural networks and heuristic methods. They use an encoder-decoder network to generate a static saliency volume, from which scanpaths are generated through stochastic techniques. In [11], the authors proposed a deep neural network, that uses an encoder-decoder network and Long Short Term Memory (LSTM) layers to generate scanpaths, combined with a adversarial training by Generative Adversarial Network (GAN) architecture.

In this paper, we introduce a new framework to predict visual attention of omnidirectional images. The proposed architecture, called SalyPath360, allows the simultaneous prediction of saliency map and a corresponding scanpath for a given stimulus.

The main contributions of this paper are as follows:

• Presenting a Neural Network that predicts saliency map and scanpath for 360° in simultaneous and parallel manner.
• Unlike existing methods, scanpaths are predicted from the refined internal features of a proven saliency prediction model.
• Improving the saliency prediction by using a joint probability mixture between the saliency map predicted by the network and saliency map constructed from the predicted scanpath.
• Use the SoftArgMax function to predict head scanpaths, and train the network seamlessly.

The rest of this paper is organised as follows: Section 2 provides a detailed description of the proposed approach, while Section 3 compares the performance of our approach against state-of-the-art methods. Finally, we conclude our study and discuss the possibility of future improvements in Section 4.

2. SalyPath360 Framework

In this section, we describe in detail the proposed framework (see Fig. 1) which is mainly composed of an encoder-decoder network augmented by a spatial attention module and an auxiliary network that takes the intermediate features at the bottleneck of the encoder-decoder network to generate the corresponding scanpath. In addition, the primary saliency map predicted by the encoder-decoder architecture and the saliency map derived from the predicted scanpath are combined to generate a more representative saliency map. The architecture of each network as well as the attention module used and the merging process applied are described below.

2.1. Encoder-Decoder Network for Saliency Prediction

The encoder-decoder network is used to predict the primary saliency map of 360° images and generate high level intermediate features exploited by the auxiliary network. Inspired by attention stream, the encoder and decoder are composed of four blocks of convolutional layers, interleaved respectively by a (2 × 2) max-pooling layer and a (2 × 2) up-sampling layer. At the bottleneck of our network, an attention module is employed to further refine the intermediate features used by the decoder and auxiliary networks. This module takes the representational feature maps and predicts a single channel heatmap, which captures a global representation covering the integrity of the stimulus. It also expands the receptive field of the encoder from (244 × 244) to (676 × 676). This expansion is vital to cover the large 360° images. The resulting heatmap is then used as a filter that refines the feature maps generated by the encoder as follows:

\[ X' = γ ⊗ X ⊗ \text{Att}_5(X) + X \]

where \( \text{Att}_5 \) is the spatial attention module function. \( X \) is the input feature maps given by the encoder. \( X' \) is the refined feature maps passed to the decoder and auxiliary networks. \( γ \) is a learnable parameter, and \( ⊗ \) represents the element-wise multiplication.

2.2. Auxiliary Network for Scanpath Prediction

An auxiliary network is also used to generate a scanpath for a given stimulus by mainly leveraging the encoding ability of the encoder-decoder network. It consists of 3 × 3 convolutional layers, each activated by a ReLU function. The last layer is composed of 100 features maps (i.e. 100 heatmaps), set in accordance to the number of fixation points per user of the considered dataset [24] (see Section ). Soft − Argmax (SAM) [21] is then used to estimate the coordinates of a fixation points from each feature map as follows:

\[ \text{SAM}(x) = \sum_{i,j \in \mathbb{Z}} \frac{e^{β_{i,j}}}{\sum_{i,j \in \mathbb{Z}} e^{β_{i,j}}} \frac{i}{W} \frac{j}{H} \]

where \( i, j, i', j' \) iterate over pixel coordinates. \( H, W \) represent the height and width of the feature map, respectively. \( x \) is the input feature map and \( β \) is a parameter adjusting the distribution of the softmax output.

As this SAM function is differentiable [21], it permits our model to be trained seamlessly unlike the discrete Argmax function. It also allows a sub-pixel accuracy, avoiding thus the use of upsampling layers to increase the size of the feature maps, and saving resources.

2.3. Merging Probability Distribution of Unbiased Saliency Map Prediction

Let us consider the predicted saliency map \( T \) constructed from encoder-decoder architecture (i.e. primary saliency map), and the saliency map \( S \) generated from the predicted fixation points by an adequate Gaussian kernel. We notice that \( T \) and \( S \) could be combined together to get a joint probability distribution [9], making thus a final representative saliency map strengthened by the most probable viewport center positions following:

\[ J(T,S) = \max_{r,s} \left( \alpha \left( \frac{T}{\text{max}_{r,s}} \right)^k + (1-\alpha) \left( \frac{S}{\text{max}_{r,s}} \right)^k \right)^{1/k} \]

\[ J(T,S) = \max_{r,s} \left( \alpha \left( \frac{T}{\text{max}_{r,s}} \right)^k + (1-\alpha) \left( \frac{S}{\text{max}_{r,s}} \right)^k \right)^{1/k} \]
2.4. Training

of head movement prediction using the following loss function on the Head-Eye movement datasets, was fine-tuned for our task encoder-decoder network based on ATSal [10], which was trained through different loss functions. More precisely, the unbiased formula in our context is defined as follows: $\text{NSS} = \frac{\sum_{i=1}^{N} \text{NSS}(T,S)}{N}$, where $\text{NSS}(T,S)$ is the Normalized Scanpath Saliency. $Y$ and $\hat{Y}$ represent the ground truth saliency map and the predicted saliency map respectively, while $Y_{fix}$ is the ground truth fixation map.

$J^*(T,S) = \left( \frac{1}{2} \times J(T,S) + \frac{1}{2} \times J(T,S) \times L_a(E) \right)$ (4)

where $L_a(E)$ represents the pixel-wise linear-scaling of the equator bias $E$.

The applied merging process algorithm for the set of pixels $T_{pixels}$ can be summarized in Algo. 1.

**Algorithm 1: JSalyPath (IN T,S,E OUT J*(T,S))**

**Result:** Joint probability distribution $J^*(T,S)$

$T$ Primary saliency map.

$S$ Scanpath-based heatmap.

$E$ Equator bias map.

Compute the normalized linear scaling of bias $L_a(E)$

$L = \frac{T_{pixels}}{2}$

for $(l \in L)$ do

$J(T,S)[l] \leftarrow$

$max \times \left( \alpha \left( \frac{T(l)}{\text{max}_{T,S}} \right)^k + (1 - \alpha) \left( \frac{S(l)}{\text{max}_{T,S}} \right)^k \right)^{1/k}$

$J^*(T,S)[l] \leftarrow$

$\left( \frac{1}{2} \times J(T,S)[l] + \frac{1}{2} \times J(T,S)[l] \times L_a(E)[l] \right)$

end

return $J^*(T,S)$;

Each term of this loss function was chosen for its own influence on the convergence of the network. Indeed, $KLD$ and $BCE$ functions minimize the distance between the distributions of the output and ground truth, while $NSS$ is a similarity metric for saliency and it is used as bias term, allowing the model to seize the saliency specific representations.

As the auxiliary branch aims to predict the coordinates of fixation points, the task can be seen as a regression problem. In addition, this branch relies upon the feature representations given by the encoder-decoder which is trained by the more complex loss function described-above. As such, this loss (i.e. $L_1$, see Eq. 5), has an indirect impact on the auxiliary branch during the training step. Therefore, we chose the Mean Squared Error (MSE) as a simple loss function $L_2$ to train this branch, defined as follows:

$L_2 = \frac{1}{N} \sum_{i=1}^{N} (p - \hat{p})^2$ (6)

where $p$ is the ground truth scanpath and $\hat{p}$ is the predicted scanpath, while $N$ is the number of fixation points of the corresponding scanpaths.

The 2 branches were trained consecutively by first fine-tuning the encoder-decoder network as it has a great influence on the accuracy of the auxiliary network. Then, the auxiliary network is trained from scratch while freezing the weights of the encoder in order to not affect the saliency prediction.
Table 1. Saliency prediction Comparison. Best results are highlighted in bold.

| Model                  | Auc Judd | Auc Borji | NSS | CC  | SIM | KLD |
|------------------------|----------|-----------|-----|-----|-----|-----|
| ATSal [10]             | 0.8479   | 0.8121    | 1.7516 | 0.6214 | 0.5748 | 1.1571 |
| Two-stream [28]        | 0.7931   | 0.7564    | 1.6249 | 0.5857 | 0.5857 | 0.8585 |
| SalyPath360 (Our method) | 0.8610   | 0.8199    | 1.8552 | 0.7194 | 0.6383 | 0.8405 |

Table 2. Scanpath prediction comparison. Best results are highlighted in bold.

| Model                  | Jarodzka | NSS |
|------------------------|----------|-----|
| PathGan [1]            | 0.1777   | -0.1518 |
| SaliNet [2]            | 0.2621   | 0.0834 |
| SalyPath360 (Our method) | 0.1363   | 0.2896 |

3. Experiments

In this section, we evaluate the ability of our model to predict saliency maps and scanpaths. We first describe the used dataset. Then, qualitative and quantitative results are presented and compared with state-of-the-art methods.

3.1. Dataset

Salient360! [24] dataset is one of the most used datasets for predicting saliency of omnidirectional images. It was proposed as a part of the 2018 Salient360! Challenge and it is composed of 85 omnidirectional images with their corresponding saliency maps and scanpaths. There are about 36 scanpaths per image where each scanpath has 100 fixation points represented by their coordinates and timestamp.

The dataset was split into training-testing sets without any overlap according to the same protocol used in [10]. The training set is composed of 70 images, while the test set is composed of 15 images, representing 82% and 18% of the dataset respectively. This choice of split was done in accordance with the common practices used in other papers [10] using this dataset. It is worth noting that for a fair comparison, all the compared models were evaluated using the same partition. To the best of our knowledge, Salient360! is the only publicly available dataset that provides the scanpaths in addition to the saliency maps for each 360° image.

3.2. Saliency Prediction

To evaluate the saliency prediction effectiveness of our method, we employed commonly used saliency metrics: Auc Judd, Auc Borji, NSS, CC, SIM, KLD [6]. The results are compared to state-of-the-art saliency models: Two-Stream [28] that achieved the best result on Salient360 challenge [18] and ATSal [10] that reached high prediction results Salient360 dataset. It is worth noting that for [28], we used the model published on the lead board of the Salient360! Challenge as Wuhan University. For [10], we used results provided by the authors for their still image model to calculate results.

Table 1 shows the results obtained. Best results are highlighted in bold. As can be seen, the proposed model outperforms all the compared saliency methods, including ATSal. For Auc Borji, we achieve a slight improvement over the state-of-the-art results, while for the other metrics high improvements are noted with a considerable improvement for NSS and CC. Fig. 2 shows a qualitative comparison between saliency maps predicted using our framework and state-of-the-art models as well the ground truth saliency maps. As can be seen, the saliency maps generated by our framework are closer to the ground-truth.

3.3. Scanpath Prediction

In this section, we evaluate the results obtained by our framework, regarding both the scanpath and the final saliency prediction using common metrics. More precisely, we employed a vector-based metric called Jarodzka [13] which compares the similarity between the scanpaths and the hybrid NSS [23] metric, which compares the scanpath with the ground truth saliency map. For the former, we applied the code used during the Salient360! challenge [15], while disregarding the temporal element which is not predicted by our framework. We also compare the performance of our framework with state-of-the-art models: PathGan [1] and SaliNet [2].

In this section, we evaluate the ability of our model to predict scanpaths and saliency maps. We first describe the used dataset. Then, qualitative and quantitative results are presented and compared with state-of-the-art methods.

3.1. Dataset

Salient360! [24] dataset is one of the most used datasets for predicting saliency of omnidirectional images. It was proposed as a part of the 2018 Salient360! Challenge and it is composed of 85 omnidirectional images with their corresponding saliency maps and scanpaths. There are about 36 scanpaths per image where each scanpath has 100 fixation points represented by their coordinates and timestamp.

The dataset was split into training-testing sets without any overlap according to the same protocol used in [10]. The training set is composed of 70 images, while the test set is composed of 15 images, representing 82% and 18% of the dataset respectively. This choice of split was done in accordance with the common practices used in other papers [10] using this dataset. It is worth noting that for a fair comparison, all the compared models were evaluated using the same partition. To the best of our knowledge, Salient360! is the only publicly available dataset that provides the scanpaths in addition to the saliency maps for each 360° image.

The results are compared to state-of-the-art saliency models: Two-Stream [28] that achieved the best result on Salient360 challenge [18] and ATSal [10] that reached high prediction results Salient360 dataset. It is worth noting that for [28], we used the model published on the lead board of the Salient360! Challenge as Wuhan University. For [10], we used results provided by the authors for their still image model to calculate results.

Table 1 shows the results obtained. Best results are highlighted in bold. As can be seen, the proposed model achieves the best results on Jarodzka and NSS. PathGan obtains the second best results on Jarodzka, while SaliNet achieves better results on NSS metric, outperforming PathGan.

A one way Analysis of Variance (ANOVA) test between groups is also applied to show if the differences between the distributions of the Jarodzka values obtained for each compared method are statistically significant. Fig. 3 shows the corresponding boxplots where the middle red mark represents the median value, and the contours of the box are the 25th and 75th percentiles. The extremities of the whiskers correspond to the minimum and maximum values without considering the outliers. The outliers are represented by red crosses and correspond to data points which are further than two or three standard deviations. As can be seen, the distributions are quite different with the smaller median value and data scatter (i.e. standard deviation), followed by those of PathGan. We then computed the p-value between the proposed method and each of the compared method (i.e. SaliNet and PathGan). The p-values were lower than the significance level (i.e. 0.05) in both cases, indicating that the differences between the distributions of the compared method are statistically significant. Fig. 3 shows qualitative comparison of the predicted and ground-truth saliency maps as well as those obtained by the encoder-decoder and the auxiliary networks. We also show the corresponding predicted scanpath using SalyPath360. As can be seen, the scanpath predicted through the proposed auxiliary network spans most of the salient regions, while maintaining the bias of equator and frontal regions, but does not visually qualify to be probable. This disparity between the quantitative results obtained through Jarodzka metric and the qualitative results shows the limitations of the metric concerning the comparison.

3.4. Impact of the joint probability merging

To assess the efficiency of the different components of our framework, we evaluated the saliency maps obtained from the predicted scanpaths, and the saliency maps obtained after the merg-
4. Conclusion

In this paper, we introduced a new framework that simultaneously predicts saliency and scanpath for omnidirectional images. The proposed model is composed of an encoder-decoder convolutional neural network for saliency prediction strengthened by an attention module, and an auxiliary network to predict the corresponding scanpath. The latter is then convolved by a Gaussian filter to derive a heatmap. A merging adaptive probability model was finally added at the end of the framework to extract a representative saliency map through the predicted saliency map and the generated scanpath-based heatmap as well as an equator bias map. In our experiments, we trained and tested the framework on the Salient360! dataset. The results were compared with state-of-the-art saliency and scanpath models, showing the effectiveness of the proposed framework for both tasks. The qualitative results also confirmed the efficiency of our model.

As perspectives, we will further improve the framework by incorporating the temporal dimension, and try to consider the inter-dependence between successive fixation points.

References

[1] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O’Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018.
[2] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O’Connor. Saltinet: Scan-path prediction on 360 degree
Figure 4. Results of the one-way Anova test for the Jaroszka values obtained for each compared method.