Scanpath and saliency prediction on 360 degree images

Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, Noel E. O'Connor

PII: S0923-5965(18)30620-9
DOI: https://doi.org/10.1016/j.image.2018.06.006
Reference: IMAGE 15402

To appear in: *Signal Processing: Image Communication*

Received date : 21 September 2017
Revised date : 13 June 2018
Accepted date : 13 June 2018

Please cite this article as: M. Assens, X. Giro-i-Nieto, K. McGuinness, N.E. O'Connor, Scanpath and saliency prediction on 360 degree images, *Signal Processing: Image Communication* (2018), https://doi.org/10.1016/j.image.2018.06.006

Scanpath and Saliency Prediction on 360 Degree Images

Marc Assens[1]and Xavier Giro-i-Nieto

Image Processing Group
Universitat Politecnica de Catalunya (UPC)
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

Kevin McGuinness and Noel E. O'Connor

Insight Center for Data Analytics
Dublin City University
Dublin, Ireland
kevin.mcguinness@insight-centre.org

Abstract

We introduce deep neural networks for scanpath and saliency prediction trained on 360-degree images. The scanpath prediction model called SaltiNet is based on a temporal-aware novel representation of saliency information named the saliency volume. The first part of the network consists of a model trained to generate saliency volumes, whose parameters are fit by back-propagation using a binary cross entropy (BCE) loss over downsampled versions of the saliency volumes. Sampling strategies over these volumes are used to generate scanpaths over the 360-degree images. Our experiments show the advantages of using saliency volumes, and how they can be used for related tasks. We also show how a similar architecture achieves state-of-the-art performance for the related task of saliency map prediction. Our source code and trained models available at https://github.com/massens/saliency-360salient-2017.

Keywords:
Deep learning, machine learning, saliency, scanpath, visual attention

[1] Work developed while Marc Assens was a visiting student at Insight Center for Data Analytics.

1. Motivation

Visual saliency prediction is a field in computer vision that aims to estimate the areas of an image that attract the attention of humans. This information can provide important clues to human image understanding. The data collected for this purpose are fixation points in an image, produced by a human observer that explores the image for a few seconds, and are traditionally captured with eye-trackers [1], mouse clicks [2], and webcams [3]. The fixations are usually aggregated and represented with a saliency map, a single channel image obtained by convolving a Gaussian kernel with each fixation. The result is a gray-scale heatmap that represents the probability of each pixel in an image being fixated by a human, and it is usually used as a soft-attention guide for other computer vision tasks.

Traditionally, saliency maps have only described fixation information with respect to the spatial layout of an image. This type of representations only encode the probability of each image pixel capture the visual attention of the user, but with no information regarding the order in which these pixels may be scanned or the duration of the fixation. Recent studies have raised the need for a representation that is also temporal-aware [4]. We address the temporal challenge for the particular case of 360° images, which contain the complete scene around the capture point and allow the viewer to choose the observation angle. Predicting the pattern that humans follow in 360° images is a topic of special interest for VR/AR applications, as it facilitates an efficient encoding and rendering on the display devices.

The main contributions of this paper are the following:

- the introduction of *saliency volumes* to capture the temporal nature of eye-gaze scan-paths;
- the SaltiNet architecture to generate scan-paths from a deep neural network that predicts saliency volumes and a sampling strategy over them;
- this work has been awarded as the best scanpath solution at the Salient360!

challenge from the IEEE International Conference on Multimedia and Expo (ICME) 2017 [5].

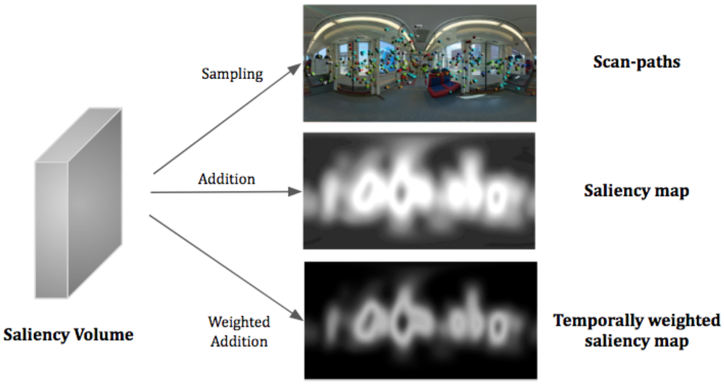


Figure 1: Scan-paths, saliency maps and temporally weighted saliency maps can be generated from a saliency volume.

This paper is structured as follows. Section 2 reviews the related literature in saliency prediction for eye fixations and presents our work with respect to them. Section 3.1 presents the whole architecture of the system, and Section 5 describes how the deep neural network was trained. Section 6 describes the experiments and results to assess the performance of the model, while Section 7 draws the conclusions and future work.

2. Related Work

2.1. Saliency prediction

The first models for saliency prediction were biologically inspired and based on a bottom-up computational model that extracted low-level visual features such as intensity, color, orientation, texture, and motion at multiple scales. Itti et al. [6] proposed a model that combines multiscale low-level features to create a saliency map. Harel et al. [7] presented a graph-based alternative that starts from low-level feature maps and creates Markov chains over various image maps, treating the equilibrium distribution over map locations as activation and saliency values.

Although these models performed reasonably well qualitatively, they had limited practical use because they frequently did not match actual human saccades from eye-tracking data. More recent research has revealed that humans not only base their attention on low-level features, but also on high-level semantics [4] (e.g., faces, humans, cars, etc.). Judd et al. introduced in [8] an approach that used low, mid, and high-level image features to define salient locations. These features where used in combination with a linear support vector machine to train a saliency model. Borji [9] also combined low-level features with top-down cognitive visual features and learned a direct mapping to eye fixations using Regression, SVM, and AdaBoost classifiers.

Recently, the field of saliency prediction has made great progress due to advance of deep learning and its applications on the task of image classification [10] [11]. The advances suggest that these models are able to capture high-level features. As noted in [4], in March of 2016 there where six deep learning models among the top 10 results in the MIT300 saliency Benchmark [12].

The enormous amount of training data necessary to train these networks makes them difficult to train directly for saliency prediction. With the objective of allowing saliency models to capture high-level features, some authors have adapted well-known models with good performance in the task of image recognition. DeepGaze [13] achieved state-of-the-art performance by reusing the well-known AlexNet [10] pretrained on ImageNet [14] with a network on top that reads activations from the different layers of AlexNet. The output of the network is then blurred, center biased, and converted to a probability distribution using a softmax. A second version called DeepGaze 2 [15] used features from VGG-19 [16] trained for image recognition. In this case, they did not fine-tune the network. Rather, some readout layers were trained on top of the VGG features to predict saliency with the SALICON dataset [2]. This results corroborated the idea that deep features trained on object recognition provide a versatile feature space for performing related visual tasks. A complete new architecture designed and trained for saliency prediction was proposed in [17], but the same work also observed the benefits of using deeper pre-trained models for image classification as a basis. Other advances in deep learning such as generative adversarial training (GANs) and attentive mechanisms have also been applied to saliency prediction: SalGAN [18] is a deep network for saliency prediction that measured the gain in performance when using a universal adversarial training in opposite to optimizing for a specific loss function. The Saliency Attentive Model (SAM) [19] includes a Convolutional LSTM that focuses on the most salient regions of the image to iteratively refine the predicted saliency map.

In [20], Torralba et al. studied how the scene modules visual attention and discovered that the same objects receive different attention depending on the scene where they appear (i.e. pedestrians are the most salient object in only 10% of the outdoor scene images, being less salient than many other objects. Tables and chairs are among the most salient objects in indoor scenes). With this insight, Liu et al. proposed DSCLRCN [21], a model based on CNNs that also incorporates global context and scene context using RNNs. Their experiments have obtained outstanding results in the MIT Saliency Benchmark.

Recently, there has been interest in finding appropriate loss functions. Huang et al. [22] made an interesting contribution by introducing loss functions based on metrics that are differentiable, such as NSS, CC, SIM, and KL divergence to train a network (see [23] and [24]).

Finally, the field of saliency prediction on omni directional images has received interested in the last years due to its applications in Virtual Reality technologies [25][26][27][28][29].

2.2. Scanpath prediction

Unlike the related task of saliency map prediction, there has not been much progress in the task of scanpath prediction over the last years. Cerf et al. [30] discovered that observers, even when not instructed to look for anything particular, fixate on a human face with a probability of over 80% within their first two fixations. Furthermore, they exhibit more similar scanpaths when faces are present. Recently, Hu et al. [31] have introduced a model capable of selecting relevant areas of a 360° video and deciding in which direction should a human observer look at each frame. An object detector is used to propose candidate objects of interest and a RNN selects the main object at each frame. d

3. SaltiNet: Scanpath prediction model

3.1. Architecture

The central element in the architecture of SaltiNet is a deep convolutional neural network (DCNN) that predicts a saliency volume for a given input image. This section provides detail on the structure of the network, the loss function, and the strategy used to generate scan-paths from saliency volumes.

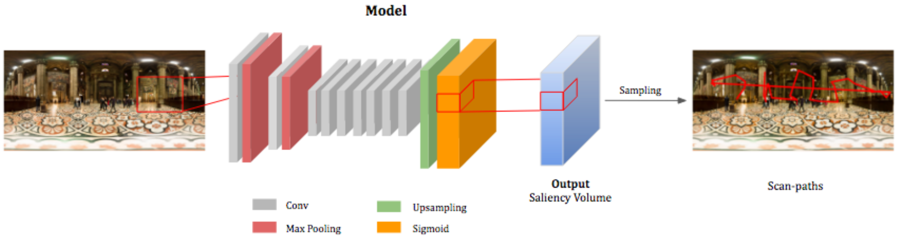


Figure 2: Overall architecture of the proposed scanpath estimation system.

3.1.1. Saliency Volumes

Saliency volumes aim to be a suitable representation of spatial and temporal saliency information for images. They have three axes that represent the width and height of the image, and the temporal dimension.

Saliency volumes are generated from information already available in current fixation datasets. First, the timestamps of the fixations are quantized. The length of the time axis is determined by the longest timestamp and the quantization step. Second, a binary volume is created by placing a '1' at fixation points and a '0' on the remaining positions. Third, a multivariate Gaussian kernel is convolved with the volume to generate the saliency volume. The values of each temporal slice are normalized, converting the slice into a probability map that represents the probability of each pixel being fixated by a user at each timestep.

Figure 1 shows how saliency volumes are a meta-representation of saliency information and how other saliency representations can be extracted from them. Saliency maps can be generated by performing an addition operation across all the temporal slices of the volume, and normalizing the values to ensure they add to one. A similar representation is the *temporally weighted saliency map*, which can be generated by performing a weighted addition operation of all the temporal slices. Finally, scan-paths can also be extracted by sampling fixation points from the temporal slices. Sampling strategies that aim to generate realistic scan-paths are discussed in Section 6.3.

3.1.2. Convolutional Neural Network

We propose a convolutional neural network (CNN) that adapts the filters learned to predict flat saliency maps to predict saliency volumes. Figure 2 illustrates the architecture of the convolutional neural network, composed of 10 layers and a total of 25.8 million parameters. Each convolutional layer is followed by a rectified linear unit non-linearity (ReLU). Excluding the last layer, the architecture follows the proposal of SalNet [17], whose first three layers are initialized from the VGG-16 model [32] trained for image classification.

Our network was designed considering the amount of training data available. Different strategies where introduced to prevent overfitting. First, the model was previously trained on the similar task of saliency map prediction, and the obtained weights were fine-tuned for the task of saliency volume prediction. Second, the input images were resized to 300 × 600, a much smaller dimension than their original size of 3000 × 6000. The last layer of the network outputs a volume of size 12 × 300 × 600, with three axis that represent time, height, and width of the image.

3.1.3. Scan-path sampling

We take a stochastic approach to scan-path sampling[2]. The generation of scan-paths from the saliency volumes requires determining: 1) number of fixations of each scan-path; 2) the duration in seconds of each fixation; and 3) the location of each fixation point. The first two values were sampled from their probability distributions learned from the training data. The location of each fixation point was also generated by sampling, this time from the corresponding temporal slice from the predicted saliency volume. Different strategies were explored for this purpose, presented together with their performance in Section 6.

4. A model for saliency map prediction

After the development of the scanpath prediction model, we also explored how a similar model based on a deep convolutional neural network could be used for the task of saliency map prediction on 360-degree images.

4.1. Architecture

The architecture of this model is similar to the one used for the previous scanpath prediction model, the main difference being the number of channels in the last convolutional layer. While the previous model featured 12 different filters in the last convolutional layer, this model only has a single filter. This results in a 2-dimensional network output suitable size for representing saliency maps.

The network is first initialized with parameters that were trained to predict saliency maps on normal images, and adapts them for the prediction of saliency maps on 360-degree images on the equirectangular space. The layers were initialized from the SalNet model [17], whose first three layers are in turn

[2] We also experimented with using an LSTM to directly predict scan-paths from the training data. However, we found that this resulted in the model regressing to the image center [33]. Future work will consider using adversarial training to address this.

initialized from the VGG-16 model [32] trained for image classification. The output of the network has a size of [300 × 600].

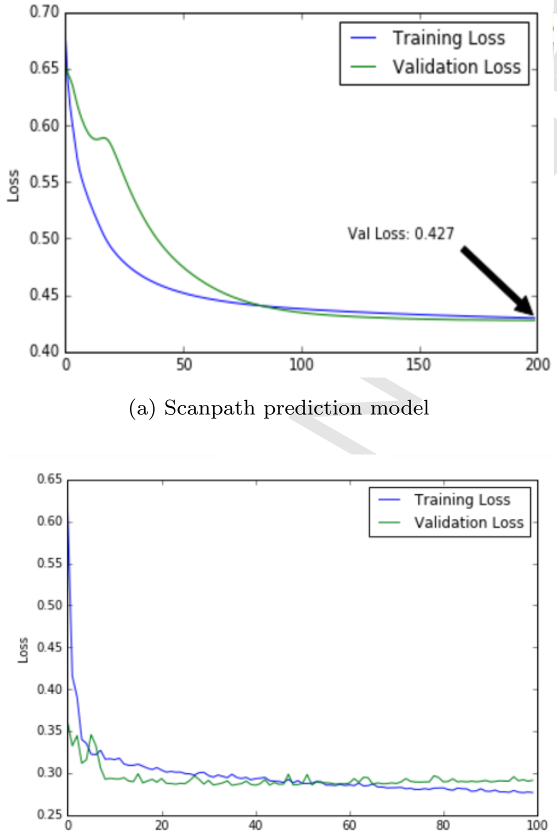
5. Training

The training procedure of the two described models is very similar. We trained the CNNs over 36 images of the 40 training images from the Salient360 dataset [5], leaving aside 4 images for validation. We normalized the values of the saliency volumes and saliency maps to be in the interval of [0, 1]. Both the input images and the output activation volumes were downsampled to 600×300 prior to training. The saliency volumes were generated from fixations by convolving with a multivariate Gaussian kernel with bandwidths $\{4, 20, 20\}$ (time, height, width). The 2D saliency maps used for training are those provided by the dataset.

The CNNs were trained using stochastic gradient descent with cross entropy loss using a batch size of 1 image for 90 epochs. The binary cross entropy loss is defined as L_{BCE} in Eq.1, where S_j and $\hat{S}_j$ correspond to the ground truth and predicted values of the saliency map.

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{j=1}^{N} S_j \log(\hat{S}_j) + (1 - S_j)\log(1 - \hat{S}_j). \tag{1}$$

During training, results on the validation set were tracked to monitor convergence and overfitting problems. The L^2 weight regularizer (weight decay) was used to avoid overfitting. Our networks took approximately two hours to train on a NVIDIA GTX Titan X GPU using the Keras framework with the Theano backend. The learning rate was set to $\alpha = 0.001$ for the scanpath prediction model, and $\alpha = 0.01$ for the saliency map prediction model. Figure 3 shows the learning curves of both models.



(a) Scanpath prediction model

(b) Saliency map prediction model

Figure 3: Training curves obtained for both models (binary cross entropy loss).

6. Experiments

6.1. Datasets

Due to the small size of the training dataset, we performed transfer learning to initialize the weights of the networks using related tasks. First, the network was trained to predict saliency maps using the SALICON dataset [22] using the same architecture of SalNet [17]. Then, the network was trained to predict saliency volumes generated from the iSUN dataset [34] that contains 6000 training images. The network was fine-tuned using the 60 images of the dataset of head and eye movements provided by the University of Nantes [35]. This dataset was acquired based on the images displayed on the head mounted display (HMD) Oculus-DK2. Eye gaze data was captured from a Sensomotoric Instruments (SMI) sensor in the HMD, which transmitted eye-tracking data binocularly at 60Hz. There were 40-42 observers, who could freely observe the scene with no task instructed. Each 360 images were shown for 25 seconds and there was a 5 second gray screen between two images.

6.2. Metrics

The similarity metric used for scanpath evaluation is a variation of the Jarodzka algorithm [36] proposed by the authors of the 360 saliency dataset [35]. The toolbox and evaluation tools can be found in [37]. The standard similarity criteria was slightly modified to use equirectangular distances in 360 instead of Euclidean distances. The generated and ground truth scan-paths are matched 1 to 1 using the Hungarian algorithm to obtain the minimum possible final cost. The presented results compare the similarity of 40 generated scan-paths with the scan-paths in the ground truth.

The evaluation of saliency map prediction has received the attention of several researchers, and there are different proposed approaches. Our experiments consider several of these, in a similar way to the MIT saliency benchmark [12].

6.3. Sampling strategies

Figure 4 shows the distribution of the number of fixations and the duration of each fixations for the training set. During scan path generation, we sample the number of fixations and their duration from these empirical distributions.

Regarding the spatial location of the fixation points, three different strategies were explored. The simplest approach (1) consists of taking one fixation for each temporal slice of the saliency volume. Through qualitative observation we noticed that scan-paths generated in this way were unrealistic, as the probability of each fixation is not conditioned on previous fixations. A more elaborated sampling strategy (2) consists of forcing fixations to be closer to their respective previous fixation. This is accomplished by multiplying a temporal slice (probability map) of the saliency volume with a Gaussian kernel centered at the previous fixation point. This suppresses the probability of positions that are far from the previous fixation point. The third sampling strategy (3) we assessed consisted of suppressing the area around all previous fixations using Gaussian kernels. As shown in Table 1, we found that the best performing model was the one using sampling strategy (2).

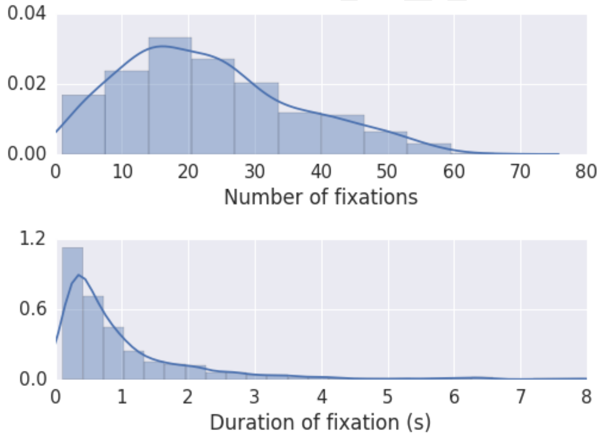


Figure 4: Empirical distributions of the number of fixations per scan-paths (top) and duration of each fixation (bottom).

6.4. Results

Scan-path prediction evaluation has received attention lately and it is a very active field of research [38][36].

Table 1 presents the impact of different sampling strategies over the saliency volume. We have compared our results with the accuracy that would be obtained by a model that outputs random fixations, and a model that outputs the ground truth fixations.

Table 2 compares our scanpath prediction model with two other solutions presented at the Salient360! Challenge [5] held at the 2017 IEEE ICME conference in Hong Kong. These figures were provided by the organizers of the

	Jarodzka↓
Random scan-paths	4.94
(1) Naive sampling strategy	3.45
(3) Avoiding fixating on same places	2.82
(2) Limiting distance between fixations	**2.27**
Sampling ground truth saliency map	1.89
Sampling ground truth saliency volume	1.79
Ground truth scan-paths	1.2e-8

Table 1: Comparison between the three considered spatial sampling strategies. Lower values are better.

	Jarodzka↓
SaltiNet (Ours)	**2.8697**
SJTU [26]	4.6565
Wuhan University [39]	5.9517

Table 2: Comparison between three submissions to the Salient360! Challenge. Lower values are better.

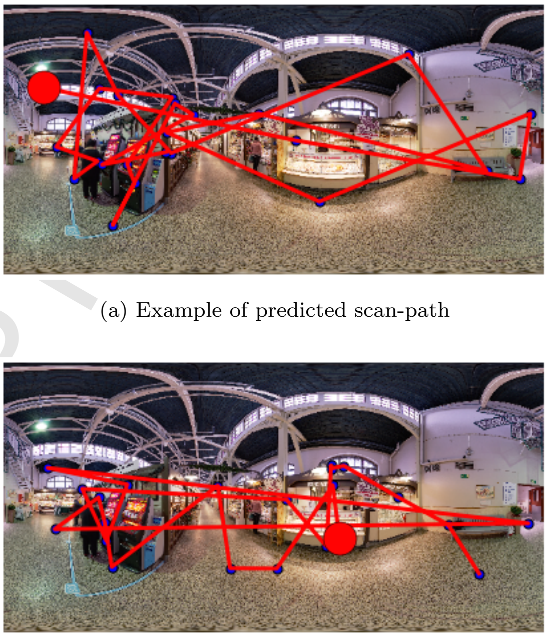
challenge. Results clearly indicate the superior performance of our system with respect to the two other participants.

The performance of our model has also been explored from a qualitative perspective by observing the generated saliency volumes and scan-paths. Figure 5 compares a generated scan-path with a ground truth scan-path. Figure 6 shows two examples of ground truth and generated saliency volumes.

Table 3 compares the performance of our saliency map prediction model with other solutions presented at the Salient360! Challenge. The results demonstrate the superior performance of our system for two of the metrics.

	KL↓	CC↑	NSS↑	ROC↑
Our Model	**0.1954**	**0.8471**	0.7785	0.6819
TU Munich, Germany [40]	0.4489	0.5786	0.8052	0.7259
SJTU, China [26]	0.4805	0.5324	0.9180	0.7347
Wuhan University, China [39]	0.5082	0.5383	**0.9358**	0.7363
TU Munich, Germany [40]	0.5008	0.5535	0.9153	**0.7467**
Zhejiang University, China [41]	0.6980	0.5270	0.8505	0.7140
Trinity College, Ireland [42]	0.4865	0.5361	0.7574	0.7019
University of Science and Technology, China [5]	2.0171	0.5073	0.9175	0.6946

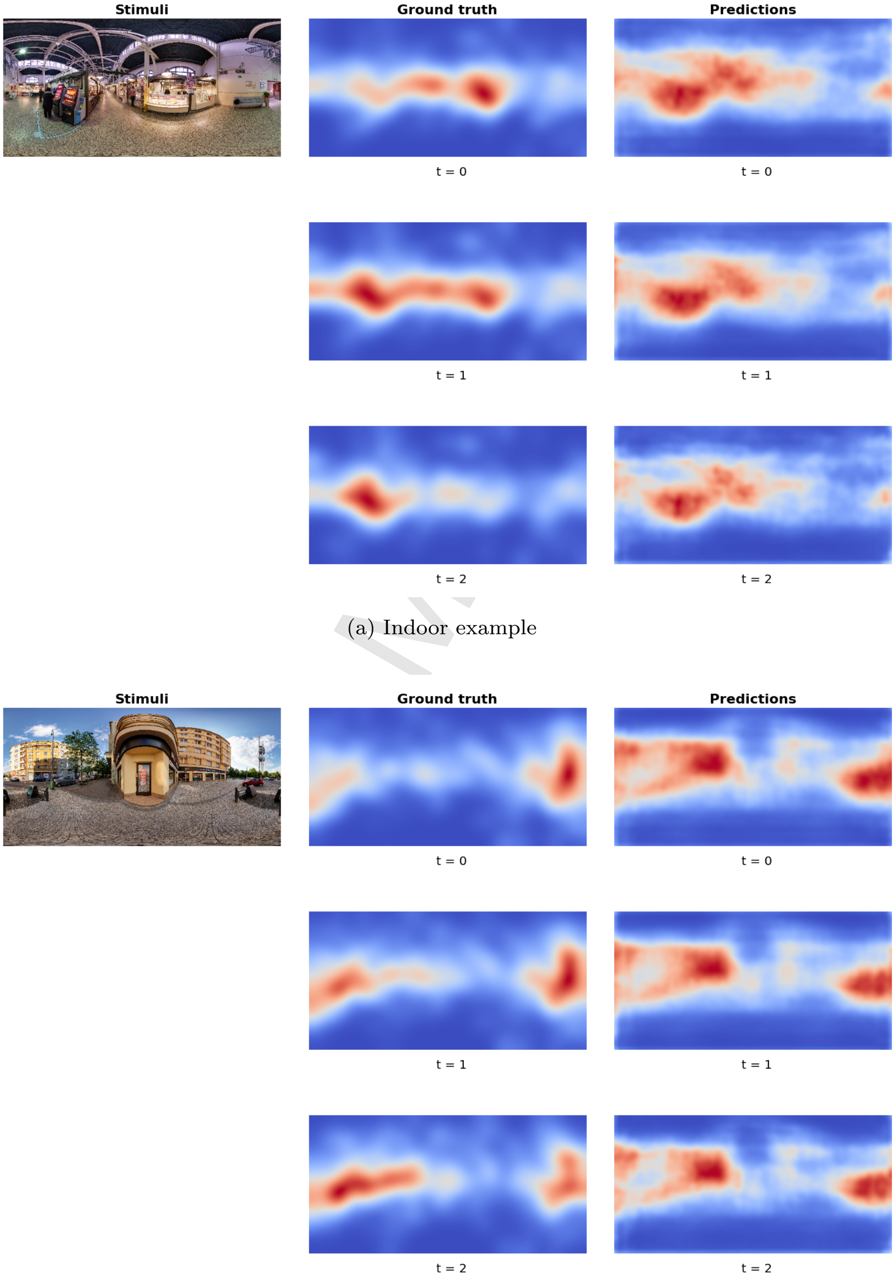
Table 3: Comparison between the submissions to the Salient360! Challenge in the saliency map prediction track.



(a) Example of predicted scan-path

(b) Example of ground truth scan-path

Figure 5: The top image shows a predicted scanpath, sampled from a predicted saliency volume. The image at the bottom shows a single ground truth scanpath.



(a) Indoor example

(b) Outdoor example

Figure 6: The images above show the predicted and ground truth saliency volumes for a given stimulus. For each saliency volume, three temporal slices are shown.

7. Conclusions

This work has presented SaltiNet, a model capable of predicting scan-paths on 360° images. This model won a performance award at the Salient360! challenge from the IEEE International Conference on Multimedia and Expo (ICME) 2017 [5]. We have also introduced a novel temporal-aware saliency representation that is able to generate other standard representations such as scanpaths, saliency maps, or temporally weighted saliency maps. Our experiments show that it is possible to obtain realistic scanpaths by sampling from saliency volumes, and the accuracy greatly depends on the sampling strategy.

We have also found the following limitations to the generation of scanpaths from saliency volumes: 1) the probability of a fixation is not conditioned to previous fixations; 2) the length of the scanpaths and the duration of each fixation are treated as independent random variables. We have tried to address the first problem by using more complex sampling strategies. Nevertheless, this three parameters are not independently distributed and therefore our model is not able to accurately represent this relationship. Future work will focus on training a fully end-to-end neural network capable of prediction the scan-paths without requiring a sampling module.

Finally, we used a very similar architecture for the related task of saliency map prediction, improving the state-of-the-art on two of the metrics. Our results can be reproduced with the source code and trained models available at `https://github.com/massens/saliency-360salient-2017`.

8. Acknowledgments

The Image Processing Group at UPC is supported by the project TEC2013-43935-R and TEC2016-75976-R, funded by the Spanish Ministerio de Economia y Competitividad and the European Regional Development Fund (ERDF). The Image Processing Group at UPC is a SGR14 Consolidated Research Group recognized and sponsored by the Catalan Government (Generalitat de Catalunya) through its AGAUR office. This material is based upon works supported by

Science Foundation Ireland under Grant No 15/SIRG/3283. We gratefully acknowledge the support of NVIDIA Corporation for the donation of GPUs used in this work.

References

[1] N. Wilming, S. Onat, J. P. Ossandón, A. Açık, T. C. Kietzmann, K. Kaspar, R. R. Gameiro, A. Vormberg, P. König, An extensive dataset of eye movements during viewing of complex images, Scientific Data 4 (2017) 160126.

[2] M. Jiang, S. Huang, J. Duan, Q. Zhao, Salicon: Saliency in context, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, A. Torralba, Eye tracking for everyone, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, F. Durand, Where should saliency models look next?, in: European Conference on Computer Vision, Springer, 2016, pp. 809–824.

[5] T. University of Nantes, Salient360: Visual attention modeling for 360 images grand challenge (2017).
URL http://www.icme2017.org/grand-challenges/

[6] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on pattern analysis and machine intelligence 20 (11) (1998) 1254–1259.

[7] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in neural information processing systems, 2007, pp. 545–552.

[8] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Computer Vision, 2009 IEEE 12th international conference on, IEEE, 2009, pp. 2106–2113.

[9] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 438–445.

[10] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[11] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in neural information processing systems, 2014, pp. 487–495.

[12] Z. Bylinskii, T. Judd, A. Ali Borji, L. Itti, F. Durand, A. Oliva, A. Torralba, MIT saliency benchmark, http://saliency.mit.edu/.

[13] M. Kümmerer, L. Theis, M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet, arXiv preprint arXiv:1411.1045.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[15] M. Kümmerer, T. S. Wallis, M. Bethge, DeepGaze II: Reading fixations from deep features trained on object recognition, ArXiv preprint:1610.01563.

[16] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[17] J. Pan, E. Sayrol, X. Giró-i Nieto, K. McGuinness, N. E. O'Connor, Shallow and deep convolutional networks for saliency prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[18] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, arXiv preprint arXiv:1701.01081.

[19] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, arXiv preprint arXiv:1611.09571.

[20] A. Torralba, A. Oliva, M. S. Castelhano, J. M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search., Psychological review 113 (4) (2006) 766.

[21] N. Liu, J. Han, A deep spatial contextual long-term recurrent convolutional network for saliency detection, arXiv preprint arXiv:1610.01708.

[22] X. Huang, C. Shen, X. Boix, Q. Zhao, Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks, in: IEEE International Conference on Computer Vision (ICCV), 2015.

[23] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, T. Dutoit, Saliency and human fixations: State-of-the-art and study of comparison metrics, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 1153–1160.

[24] M. Kümmerer, T. S. Wallis, M. Bethge, Information-theoretic model comparison unifies saliency metrics, Proceedings of the National Academy of Sciences 112 (52) (2015) 16054–16059.

[25] Y. Rai, P. Le Callet, P. Guillotel, Which saliency weighting for omni directional image quality assessment?, in: Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, IEEE, 2017, pp. 1–6.

[26] Y. Zhu, G. Zhai, X. Min, The prediction of head and eye movement for 360 degree images, Signal Processing: Image Communication.

[27] J. Ling, K. Zhang, Y. Zhang, D. Yang, Z. Chen, A saliency prediction model on 360 degree images using color dictionary based sparse representation, Signal Processing: Image Communication.

[28] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, G. Wetzstein, Saliency in vr: How do people explore virtual environments?, arXiv preprint arXiv:1612.04335.

[29] A. De Abreu, C. Ozcinar, A. Smolic, Look around you: Saliency maps for omnidirectional images in vr applications, in: Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, IEEE, 2017, pp. 1–6.

[30] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, in: Advances in neural information processing systems, 2008, pp. 241–248.

[31] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, M. Sun, Deep 360 pilot: Learning a deep agent for piloting through 360 degree sports video, arXiv preprint arXiv:1705.01759.

[32] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531.

[33] M. Mathieu, C. Couprie, Y. LeCun, Deep multi-scale video prediction beyond mean square error, arXiv preprint arXiv:1511.05440.

[34] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, J. Xiao, Turkergaze: Crowdsourcing saliency with webcam based eye tracking, arXiv preprint arXiv:1504.06755.

[35] Y. Rai, J. Gutiérrez, P. Le Callet, A dataset of head and eye movements for 360 degree images, in: Proceedings of the 8th ACM on Multimedia Systems Conference, ACM, 2017, pp. 205–210.

[36] H. Jarodzka, K. Holmqvist, M. Nyström, A vector-based, multidimensional scanpath similarity measure, in: Proceedings of the 2010 symposium on eye-tracking research & applications, ACM, 2010, pp. 211–218.

[37] J. Gutiérrez, E. David, Y. Rai, P. Le Callet, Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360° still images, Signal Processing: Image Communication.

[38] O. Le Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior research methods 45 (1) (2013) 251–266.

[39] Y. Fang, X. Zhang, A novel superpixel-based saliency detection model for 360-degree images, Signal Processing: Image Communication.

[40] M. Startsev, M. Dorr, 360-aware saliency estimation with conventional image saliency predictors, Signal Processing: Image Communication.

[41] P. Lebreton, A. Raake, Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images, Signal Processing: Image Communication.

[42] R. Monroy, S. Lutz, T. Chalasani, A. Smolic, Salnet360: Saliency maps for omni-directional images with cnn, arXiv preprint arXiv:1709.06505.

The main contributions of this paper are the following:

- the introduction of saliency volumes to capture the temporal nature of eye-gaze scan-paths;

- the SaltiNet architecture to generate scan-paths from a deep neural network that predicts saliency volumes and a sampling strategy over them;

- this work has been awarded as the best scanpath solution at the Salient360! challenge from the IEEE Inter- national Conference on Multimedia and Expo (ICME) 2017 [29].

- A similar architecture to SaltiNet suitable for saliency map prediction with state of the art performance