PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks

Marc Assens[1], Kevin McGuinness[1],
Xavier Giro-i-Nieto[2], and Noel E. O'Connor[1]

[1] Insight Centre for Data Analytic, Dublin City University. Dublin, Ireland.
kevin.mcguinness@insight-centre.org
[2] Universitat Politecnica de Catalunya. Barcelona, Catalonia/Spain.
xavier.giro@upc.edu

Abstract. We introduce PathGAN, a deep neural network for visual scanpath prediction trained on adversarial examples. A visual scanpath is defined as the sequence of fixation points over an image defined by a human observer with its gaze. PathGAN is composed of two parts, the generator and the discriminator. Both parts extract features from images using off-the-shelf networks, and train recurrent layers to generate or discriminate scanpaths accordingly. In scanpath prediction, the stochastic nature of the data makes it very difficult to generate realistic predictions using supervised learning strategies, but we adopt adversarial training as a suitable alternative. Our experiments prove how PathGAN improves the state of the art of visual scanpath prediction on the Salient360! dataset.

Keywords: saliency, scanpath, adversarial training, GAN, 360 video

1 Motivation

When a human observer looks at an image, he spends most of his time looking at specific regions [13, 1]. He starts directing his gaze at a specific point and explores the image creating a sequence of fixation points that covers the salient areas of the image. This process can be seen as a resource allocation problem; our visual system decides where to direct its attention, in which order, and how much time will be spent in each location given an image.

There are namely two ways of predicting the user attention over a scene: whether as a saliency map [8], in which a probability value of fixation is assigned to each pixel in an image, or as a scanpath prediction, in which a sequence of gaze fixations is predicted. This work focuses in the later, as it is richer in the sense that an aggregation of scanpaths would actually generate the saliency map of an image.

Scanpath prediction has been previously studied in different domains. Firstly, in a classic set up in which a viewer is observing an external scene, for example, by passively looking at a recorded image or video. [10]. Other works have focused in egocentric vision from wearables, where the user wears a camera and actively

chooses its orientation in a real life set up [3, 4, 7]. More recently, the maturity 360 degree cameras for virtual and augmented reality has introduced another domain [2], as predicting the scanpath the user's gaze will follow can provide valuable information in terms of efficient rendering or coding.

This paper explores an end-to-end solution for omni directional scanpath prediction using conditional adversarial training. We show that this framework is suitable for this task. Our results achieve state-of-the-art performance using a convolutional-recurrent architecture in the Salient360 Dataset [6].

2 Related Work

User attention and saliency prediction in 360-degree videos has been previously explored from different perspectives. Ling et al [11] propose a solution based on color features and sparse representation to predict the saliency map of the head motion. However, Rai et al. [15] point that predicting the orientation of the head mounted display is not enough to estimate the saliency over scenes, as the gaze fixation does not correspond to simply the center bias of the head orientation.This motivated the first works in scanpath prediction in 360 degree videos, all of them based on an initial prediction of a saliency map. Zhu et al. [17] compute saliency maps based on handcrafted features to later identify clusters over them that will define the fixation points. Assens et al [2] proposed SalTiNet, a deep learning approach that generates a novel three-dimensional representation of saliency maps: the *saliency volumes*. This data structure captured the temporal location of the fixation across an additional temporal axis added to the classic saliency maps. The final scanpath was generated by sampling fixation points from this saliency volumes and finally introducing a post-filtering stage. PathGAN also uses a deep neural model, but provides a fully end-to-end solution where the model directly generates a scanpath, with no need of any sampling nor post-processing.

3 Adversarial Training for Sequence Generation

PathGAN is based on the seminal work of Ian Goodfellow et al. on Generative Adversarial Networks (GANs) [5]. In an adversarial framework, two models are trained iteratively. First, the generative model G tries to capture the data distribution. Second, the discriminator model D estimates the probability that a given sample is synthesized or real. During training, G tries to maximize the probability of fooling D. This process can also be seen as if GANs learn a loss function to tell if a sample is real or fake. Generated samples that are not realistic (e.g. blurry images, or scanpaths with all the fixations in the center) will not be tolerated.

A popular variation of GANs are the Conditional Adversarial Networks (cGANs) [14], where G does not output a sample purely from a noise vector, but it is also conditioned on a given input vector. In this setting, D needs to observe the conditioning vector to decide about the nature of the sample to be

classified into synthesized or real. In our work, we adopt the cGAN paradigm to overcome the limitation reported in [2] when trying to use a RNN for visual scanpath prediction. This way, PathGAN proposes to train a RNN following an adversarial approach, in such a way that the resulting generator produces realistic and diverse scanpaths conditioned to the input image.

The overall architecture of PathGAN is depicted in Figure 1. It is composed by two deep neural networks, the generator and the discriminator, whose combined efforts aim at predicting a realistic scanpath from a given image. The model is trained following the cGAN framework to allow the predictions to be conditioned to an input image, encoded by a pre-trained convolutional neural network. This section provides details about the structure of both networks and the considered loss functions.

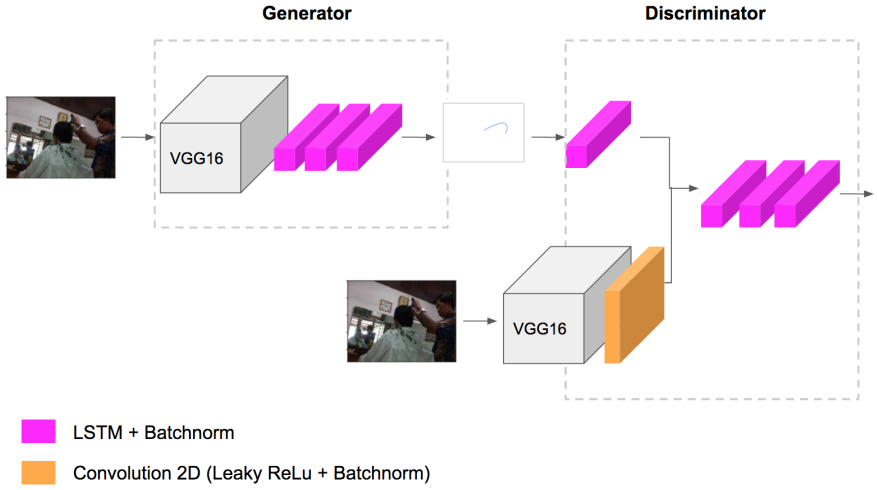

Fig. 1. Overall architecture of the proposed convolutional-recurrent model

4 Experiments

PathGAN was initially trained on the iSUN dataset [16] that contains 6,000 training images. and later fine-tuned to predict scanpaths on omni directional images using the Salient360 dataset, which contains 60 training images with data obtained from head and eye movements from the human observers.

The similarity metric used in the experiments is the Jarodzka algorithm [9]. This metric presents different advantages over other common metrics like the Levenshtein distance or correlating attention maps. In the first place, it preserves the overall shape, direction and amplitude of the saccades, the position and

duration of the fixations. Second, it provides more detailed information on the type of similarity between two vectors. This metric was used in the Salient360, scanpath prediction challenge at ICME 2017 [12]. The implementation of the metric for omni directional images was released by the University of Nantes [6]. This code was adapted to compute the Jarodzka metric for conventional images on the iSUN dataset. The ground truth and predicted scanpaths are then matched 1-to-1 using the Hungarian algorithm to obtain the obtain the minimum cost.

Table 1 compares the performance on omni directional images using the Jarodzka metric, against other solutions presented at the Salient360! Challenge [12], which took place at the IEEE ICME 2017 conference in Hong Kong. The results of the participants were calculated by the organization on a test set whose ground truth was not public at the time.

	Wuhan [11]	SJTU [17]	SaltiNet [2]	PathGAN
Jarodzka ↓	5.95	4.66	2.87	**0.74**

Table 1. Comparison with the best submissions to the ICME 2017 Salient360! Lower values are better.

Figure 2 depicts qualitative results of the generated scanpaths.

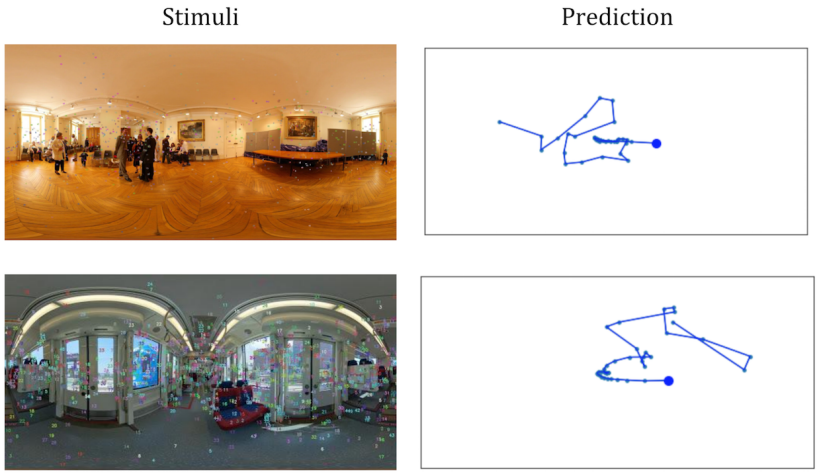

Fig. 2. Examples of predictions on the Salient360! dataset. The stimuli has the ground truth annotated.

References

1. Amor, T.A., Reis, S.D., Campos, D., Herrmann, H.J., Andrade Jr, J.S.: Persistence in eye movement during visual search. Scientific reports **6**, 20815 (2016)
2. Assens, M., Giro-i Nieto, X., McGuinness, K., OConnor, N.E.: Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW). pp. 2331–2338. IEEE (2017)
3. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Advances in neural information processing systems. pp. 241–248 (2008)
4. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: European Conference on Computer Vision. pp. 314–327. Springer (2012)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
6. Gutiérrez, J., David, E., Rai, Y., Le Callet, P.: Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360° still images. Signal Processing: Image Communication (2018)
7. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. arXiv preprint arXiv:1803.09125 (2018)
8. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence **20**(11), 1254–1259 (1998)
9. Jarodzka, H., Holmqvist, K., Nyström, M.: A vector-based, multidimensional scanpath similarity measure. In: Proceedings of the 2010 symposium on eye-tracking research & applications. pp. 211–218. ACM (2010)
10. Jiang, M., Boix, X., Roig, G., Xu, J., Van Gool, L., Zhao, Q.: Learning to predict sequences of human visual fixations. IEEE Trans. Neural Netw. Learning Syst. **27**(6), 1241–1252 (2016)
11. Ling, J., Zhang, K., Zhang, Y., Yang, D., Chen, Z.: A saliency prediction model on 360 degree images using color dictionary based sparse representation. Signal Processing: Image Communication (2018)
12. University of Nantes, T.: Salient360: Visual attention modeling for 360 images grand challenge (2017), http://www.icme2017.org/grand-challenges/
13. Porter, G., Troscianko, T., Gilchrist, I.D.: Effort during visual search and counting: Insights from pupillometry. The Quarterly Journal of Experimental Psychology (2007)
14. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
15. Rai, Y., Le Callet, P., Guillotel, P.: Which saliency weighting for omni directional image quality assessment? In: Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on. pp. 1–6. IEEE (2017)
16. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)
17. Zhu, Y., Zhai, G., Min, X.: The prediction of head and eye movement for 360 degree images. Signal Processing: Image Communication (2018)