How to Look Next? A Data-Driven Approach for Scanpath Prediction

Giuseppe Boccignone, Vittorio Cuculo(✉), and Alessandro D'Amelio

PHuSe Lab - Dipartimento di Informatica, University of Milan, Milan, Italy
{giuseppe.boccignone,vittorio.cuculo,alessandro.damelio}@unimi.it

Abstract. By and large, current visual attention models mostly rely, when considering static stimuli, on the following procedure. Given an image, a saliency map is computed, which, in turn, might serve the purpose of predicting a sequence of gaze shifts, namely a scanpath instantiating the dynamics of visual attention deployment. The temporal pattern of attention unfolding is thus confined to the scanpath generation stage, whilst salience is conceived as a static map, at best conflating a number of factors (bottom-up information, top-down, spatial biases, etc.).

In this note we propose a novel sequential scheme that consists of a three-stage processing relying on a center-bias model, a context/layout model, and an object-based model, respectively. Each stage contributes, at different times, to the sequential sampling of the final scanpath. We compare the method against classic scanpath generation that exploits state-of-the-art static saliency model. Results show that accounting for the structure of the temporal unfolding leads to gaze dynamics close to human gaze behaviour.

Keywords: Saliency model · Visual attention · Gaze deployment · Scanpath prediction

1 Introduction

Background. The unfolding of visual attention deployment in time can be captured at the *data level* by eye-tracking the observer while scrutinising for a time T a scene, either static or dynamic, under a given task or goal. Figure 1 (left panel) summarises the process.

The raw gaze trajectories can be subsequently parsed in a discrete sequence of time-stamped gaze locations or fixations $(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \cdots$, a scanpath, where the displacement from one fixation to the next might occur as a quick jump/flight (saccade) or through the smooth pursuit of a moving item in the scene. Further, by collecting the fixations of S subjects on the i-th stimulus, an attention map or heat map can be computed in the form of a 2D empirical fixation distribution map, say $\mathcal{M}_T^{D(i)}$. At the *model level*, given a stimulus and an initial gaze point, attentive eye guidance entails answering the question: *Where to Look Next?* In a nutshell, the "Where" part concerns choosing *what* to gaze at

E. Sekerinski et al. (Eds.): FM 2019 Workshops, LNCS 12232, pp. 131–145, 2020.
https://doi.org/10.1007/978-3-030-54994-7_10

- features, objects, actions - and their location; the "Next" part involves *how* we gaze at what we have chosen to gaze, that is directly affected by factors such as context [36], spatial biases [34], affect and personality [17] and crucially brings in the unfolding dynamics of gaze deployment.

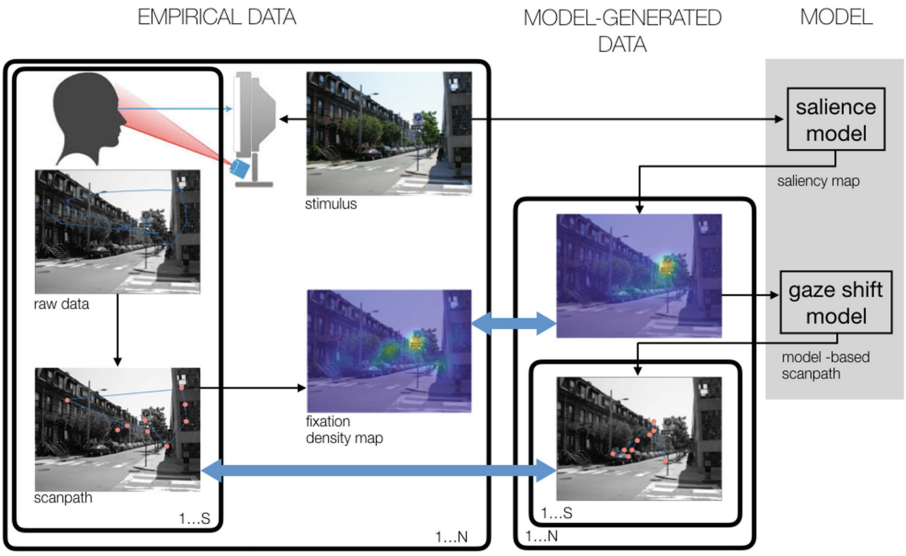


Fig. 1. Gaze data recording via eye-tracking and modelling. Given a stimulus (image **I**), the observer's gaze trajectory is sampled and recorded. Raw data are parsed and classified in fixations sequences (scanpaths). Collecting fixations from all subjects the 2D empirical fixation distribution $\mathcal{M}^D$ is estimated. On the model side, for the same stimulus a saliency map $\mathcal{S}$ is derived; if available, a gaze shift model can be exploited for sampling scanpaths based on $\mathcal{S}$. The overall model performance is routinely evaluated by comparing either the model-generated saliency map $\mathcal{S}$ with the empirical $\mathcal{M}^D$ map (light blue two-head arrows) and/or, albeit less commonly, by confronting the model-generated scanpaths $\{\widetilde{\mathbf{r}}_F(1), \widetilde{\mathbf{r}}_F(2), \cdots\}$, with the actual ones $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$.

More formally, a computational model of visual attention deployment should account for the mapping from visual data of a natural scene, say $\mathbf{I}$ (raw image data, either a static picture or a stream of images), to the scanpath

$$\mathbf{I} \mapsto \{\mathbf{r}_{F_1}, t_1; \mathbf{r}_{F_2}, t_2; \cdots\}. \tag{1}$$

When dealing with static stimuli (images) such mapping boils down to the following (cfr. Fig.1, right panel)

1. Compute a saliency map $\mathcal{S}$, i.e.,

$$\mathbf{I} \mapsto \mathcal{S}; \tag{2}$$

2. Use $\mathcal{S}$ to generate the scanpath,

$$\mathcal{S} \mapsto \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}, \tag{3}$$

where we have adopted the compact notation $(\mathbf{r}_{F_n}, t_n) = \mathbf{r}_F(n)$.

In its original formulation [21], the "saliency map" $\mathcal{S}$ is a topographic representation indicating *where* one is likely to look within the viewed scene, that is $\mathcal{S}(\mathbf{r}) \approx P(\mathbf{r} \mid \mathbf{F}(\mathbf{I}))$, where $\mathbf{F}(\mathbf{I})$ are low-level features computed from image $\mathbf{I}$. In a sense, it can be considered the modelling counterpart of the fixation density map $\mathcal{M}_T^{D(i)}$. Notice that, in recent years, computer vision efforts to achieve benchmarking performance have resulted in the heuristic addition of high-level processing capabilities to attention models, which are still referred to as salience models [9–13]. As a matter of fact, the term "saliency" now stands for any image-based prediction of which locations are likely to be fixated by subject guided by either low- or high-level cues [29].

Challenges. Despite of the original purpose behind steps 1 and 2, i.e. computing the mapping in Eq. 1, it is easily recognised by overviewing the field [9–11,33], that computational modelling of visual attention has been mainly concerned with stage 1, that is calculating salience $\mathcal{S}$. As to stage 2, it is seldom taken into account: as a matter of fact, it is surmised that $\mathcal{S}$ is *per se* predictive of human fixations. Thus, saliency models to predict where we look have gained currency for a variety of applications in computer vision, image and video processing and compression, quality assessment.

Under such circumstances, a crucial and often overlooked problem arises: saliency maps do not account for temporal dynamics. In current practice, saliency models are learned and/or evaluated by simply exploiting the fixation map on an image as "freezed" at the end of the viewing process (i.e, after having collected all fixations on stimulus along an eye-tracking session). The temporal pattern of attention unfolding, whether considered, is thus confined to the scanpath generation stage (Eq. 3), whilst salience $\mathcal{S}$ is conceived as a static map, at best simultaneously conflating a number of factors (bottom-up information, top-down, spatial biases, etc.) In simple terms, the unfolding of visual attention does not unfold.

Our Approach. In an earlier communication [7], it has been shown that the evolution of the empirical fixation density $\mathcal{M}_t^{D(i)}$ within the time interval $[t_0, T]$ from the onset of the stimulus i up to time T, provides a source of information which is richer than that derived by simply considering its cumulative distribution function $\int_{t_0}^{T} \mathcal{M}_t^{D(i)} dt$. By resorting to a simulation of scanpath generation from empirical fixation densities collected at different stages of attention unfolding, it was possible to show that:

(i) the scanpaths sampled in such way considerably differ from those generated by a static attention map;
(ii) "time-aware" scanpaths exhibit a dynamics akin to that of actual scanpaths recorded from human observers.

More precisely, those analyses [7] were based on sequentially computing, from empirical data, three different fixation density maps $\mathcal{M}_{t_k}^{D(i)}$, within the time interval $[t_0, T]$, $k = 1, 2, 3$ with $t_k < t_j$, for $k < j$, thus with time delays $D_k = t_k - t_0$. Each map was used to sample the partial scanpath related to that specific time window.

In the work presented here, we operationally take into account such temporal aspects of attention deployment (Sect. 2). In brief, we provide a "time-aware" model that addresses the three stages described above by exploiting a center bias model, a context model and an object model whose output maps are sequentially used to sample gaze shifts contributing to the final scanpath (cfr. Fig. 2, below)

We show (Sect. 3) that in such way the model-based sampling of gaze shifts, which simulates *how* human observers actually allocate visual resources onto the scene (i.e., the scanpath), departs from that achieved by classic modelling relying on a unique static saliency map (Eqs. 2 and 3), and it exhibits the features noticed in preliminary analyses based on empirical data [7].

2 A Model for Time-Aware Scanpath Generation

Recent work by Schutt *et al.* [32] has considered the temporal evolution of the fixation density in the free viewing of static scenes. They have provided evidence for a fixation dynamics which unfolds into three stages:

1. An initial orienting response towards the image center;
2. A brief exploration, which is characterized by a gradual broadening of the fixation density, the observers looking at all parts of the image they are interested in;
3. A final equilibrium state, in which the fixation density has converged, and subjects preferentially return to the same fixation locations they visited during the main exploration.

In [7] it has been shown that by estimating from eye-tracking data the empirical fixation distribution $\mathcal{M}_k^{D(i)}$ at each temporal stage described above and using it to sample a partial scanpath $\mathcal{R}t_k^{(s,i)}, k = 1, 2, 3$, eventually the "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$ more closely resembles human scanpaths than scanpaths classically obtained from the final attention map.

The main goal of this note is thus to outline a model to substantiate such results. In brief, the scheme we propose consists of a three-stage processing where the dynamics described by Schutt *et al.* [32] basically relies on: 1) a center-bias model for initial focusing; 2) a context/layout model accounting for the broad exploration to get the gist of the scene; an object-based model, to scrutinise objects that are likely to be located in such context. The output of each model is a specific map, guiding, at a that specific stage, the sequential sampling of a partial scanpath via the gaze shift model. The three-stage model is outlined at a glance in Fig. 2. The overall model dynamics can be described as follows. Given the i-th image stimulus at onset time t_0:

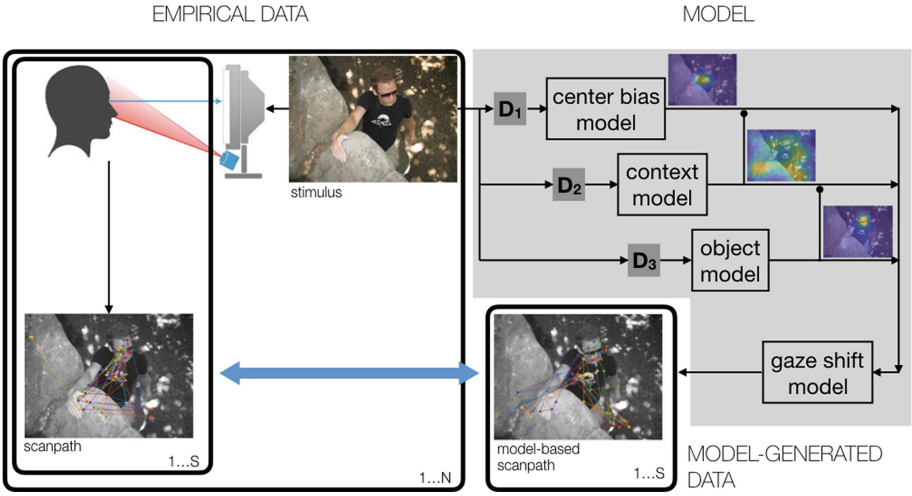


Fig. 2. The proposed three-stage model. The "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$, for each "artificial observer" s viewing the i-th stimulus, is obtained from the three partial scanpaths. These are sampled by relying on the three maps computed via the center bias, context and object models, respectively. Each model m is activated at a delay time D_m, while inhibiting the output of model $m-1$, so that the gaze model operates sequentially in time on one and only map. Empirical data collection is organised as outlined in Fig. 1. Here, the overall model performance is assessed by comparing the model-generated scanpaths $\{\widetilde{\mathbf{r}}_F(1), \widetilde{\mathbf{r}}_F(2), \cdots\}$, with the actual ones $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$.

For all stages $k = 1, 2, 3$

Step 1. At time delay D_k, compute the model-based map $\mathcal{M}_k^{(i)}$

Step 2. Based on $\mathcal{M}_k^{(i)}$, generate "subject" fixations via the gaze shift model $\mathbf{r}_F^{(s,i)}(n) = f(\mathbf{r}_F^{(s,i)}(n-1), \mathcal{M}_k^{(i)})$:

$$\mathcal{M}_k^{(i)} \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(m_{k-1}+1), \cdots, \widetilde{\mathbf{r}}_F^{(s,i)}(m_k)\} = \mathcal{R}t_k^{(s,i)}, \tag{4}$$

Eventually, collect the "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$.

For what concerns scanpath sampling, as proposed in [7], we exploit the Constrained Levy Exploration (CLE [3]) model, that has also been widely used for evaluation purposes, e.g., [24,38].

More specifically we consider the following model components to compute the maps $\mathcal{M}_k^{(i)}, k = 1, 2, 3$.

2.1 Center Bias

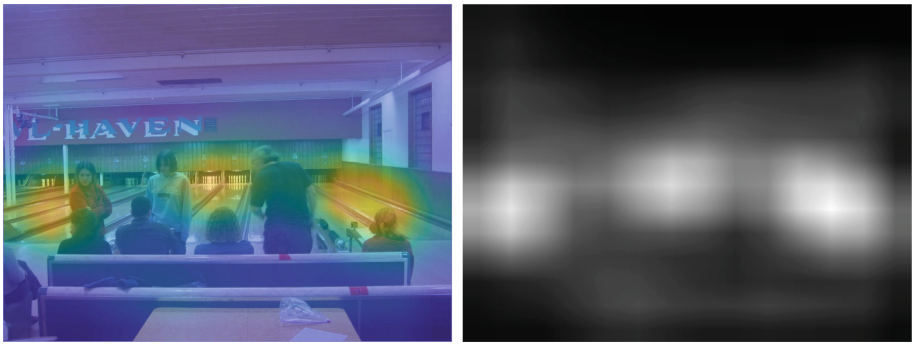
Many studies [31,37] of attentional selection in natural scenes have observed that the density of the first fixation shows a pronounced initial center bias caused by

a number of possible factors: displacement bias of an image content (known as photographer bias), motor bias (related to the experiment protocol) as well as physical preferences in orbital position. In this study the center bias is modelled with a bidimensional Gaussian function located at the screen center with variance proportional to the image size, as shown in the first column of Fig. 5.

2.2 Context Model

Behavioural experiments [30] on scene understanding demonstrated that humans are able to correctly identify the semantic category of most real-world scenes even in case of fast and blurred presentations. Therefore, objects in a scene are not needed to be identified to understand the meaning of a complex scene. The rationale presented in [30], where a formal approach to the representation of scene *gist* understanding is presented, was further developed in [42] addressing scene classification via CNNs. The models were trained on the novel Places database consisting of 10 million scene photographs labelled with environment categories. In particular, we exploited the WideResNet [39] model fine-tuned on a subset of the database consisting of 365 different scene categories. The context map, therefore, is the result of the top-1 predicted category Class Activation Map (CAM) [41]. CAM indicates the discriminative image regions used by the network to identify a particular category and, in this work, simulates the exploration phase during which observers look at those portions of the image which are supposed to convey the relevant information for the scene context understanding.

In Fig. 3 is shown an example extracted from the dataset adopted in Sect. 3, where a bowling alley is correctly identified by the network when focusing on the bowling lanes.



(a) Scene, predicted as "bowling_alley" (b) Context map

Fig. 3. Components of the context map. In (a) is shown the Class Activation Map of a scene correctly identified as "bowling alley", while in (b) the corresponding considered context map

2.3 Object Model

The last stage to the realisation of the final scanpath accounts for the convergence of fixations on relevant objects.

It is worth noting that the relevance of an object is in principle strictly related to a given task [33]. The study presented here relies on eye-tracking data collected from subjects along a free-viewing (no external task) experiment and the sub-model design reflects such scenario. However, even under free-viewing conditions, it has been shown that at least faces and text significantly capture the attention of an observer [14]. Clearly, when these kinds of object are missing, other common objects that might be present within the scene become relevant.

In order to obtain a realistic object map we exploited three different sub-frameworks implementing face detection, text detection and generic object segmentation, respectively. The output of each detector contributes, with different weight, to the final object map.

More specifically, the face detection module relies on the HR-ResNet101 network [20] that achieves state-of-the-art performance even in presence of very small faces. This extracts canonical bounding box shapes that identify the regions containing a face. An example of the face detection phase is provided in Fig. 4a.

(a) Faces (0.5) (b) Objects (0.4) (c) Text (0.1)

(d) Object map

Fig. 4. Components of the object map: (a) shows the result of the face detector module; (b) the result of the object segmentation; (c) text detection result. In brackets, the weights of each component, in terms of contribution to the final object map (d).

The generic object detection component is implemented via Mask R-CNN [18,19]. The latter capture objects in an image, while simultaneously generating a high-quality segmentation mask for each instance. The CNN is trained on the COCO dataset [27], that consists of natural images that reflect everyday scene and provides contextual information. Multiple objects in the same image are annotated with different labels, among a set of 80 possible object categories, and segmented properly. Figure 4b shows an example, where all persons present in the image, as well as traffic lights and cars are precisely identified and segmented. The text detection component is represented by a novel Progressive Scale Expansion Network (PSENet) [26], which can spot text with arbitrary shapes even in presence of closely adjacent text instances. An example of text detection result is shown in Fig. 4c.

3 Simulation

Dataset. The adopted dataset is the well-known MIT1003 [22], that consists of eye tracking data (240 Hz) recorded from $N_S = 15$ viewers during a free-viewing experiment involving 1003 natural images. The stimuli were presented at full resolution for 3 s. The raw eye tracking data were classified in fixations and saccades by adopting an acceleration threshold algorithm [22].

Evaluation. As described in Sect. 2, we generated four different maps for each image $\mathbf{I}^i$ of the dataset. Three of these are the results of the adopted sub-models: center bias, context and object. The latter is obtained by combining the outputs of the three detectors: faces, text and common objects. The first two are the most relevant cues [14] and we empirically assigned weights 0.5 and 0.4, respectively, while weighting 0.1 the object segmentation result. The final object map is later normalised to deal with possible lacks of any of the three components.

The comparison was carried out with the state-of-the-art static saliency model DeepGaze II [23]. This is based on deep neural network features pre-trained for object recognition. The model is later fine-tuned on the MIT1003 dataset and the center bias is explicitly modelled as a prior distribution that is added to the network output. The prior distribution is the result of a Gaussian kernel density estimation over all fixations from the training dataset.

All the considered saliency maps are convolved with a Gaussian kernel with $\sigma = 35$ px (corresponding to 1dva for the MIT1003 dataset). Figure 5 shows examples of the generated maps.

These were used to support the generation of $N_S = 15$ scanpaths for both the proposed and DeepGaze II approach, via the CLE gaze shift model[1] [3]. The number of fixations generated for each subject is sampled from the empirical distribution of the number of fixations performed by the human observer over each stimulus. Furthermore, in the proposed model, the switching time from the center bias map to the context map is set to 500 ms, while the permanence of the second map is equal to 1000 ms and the sampling of fixations from the object

[1] Code available at https://github.com/phuselab/CLE.

map is done for 1500 ms. In terms of delay time D_m, each model m is activated at $D_m = \{0, 500, 1000\}$ ms, while inhibiting the output of model $m - 1$.

Figure 6 shows CLE generated scanpaths, compared against the actual set of human scanpaths. The examples show how considering the context in the exploration of a scene and the precise detection of salient high-level objects, leads to scanpaths that are closer to those resulting from human gaze behaviour, than scanpaths generated via the classic saliency map. In particular, the first two rows of Fig. 6 show how the contribution of the context map reflects the human exploration of the background, rather than focusing only on faces. The third row shows an example where DeepGaze II gives high relevance to low-level features that are not salient for human observers. In the following row it can be noticed how during the exploration phase all the faces are relevant, even when these are

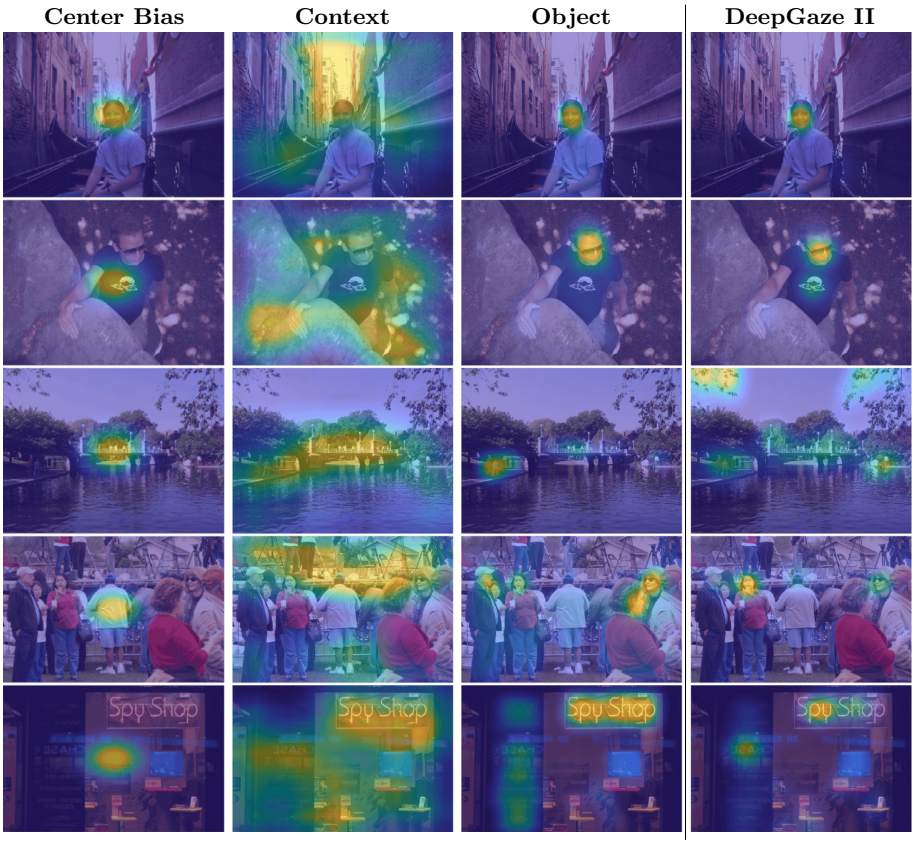


Fig. 5. Example of different maps generated for five images extracted from MIT1003 dataset. From left to right: the center bias, the context map and the object map, superimposed on the original stimulus; the saliency map resulting from saliency model DeepGaze II.

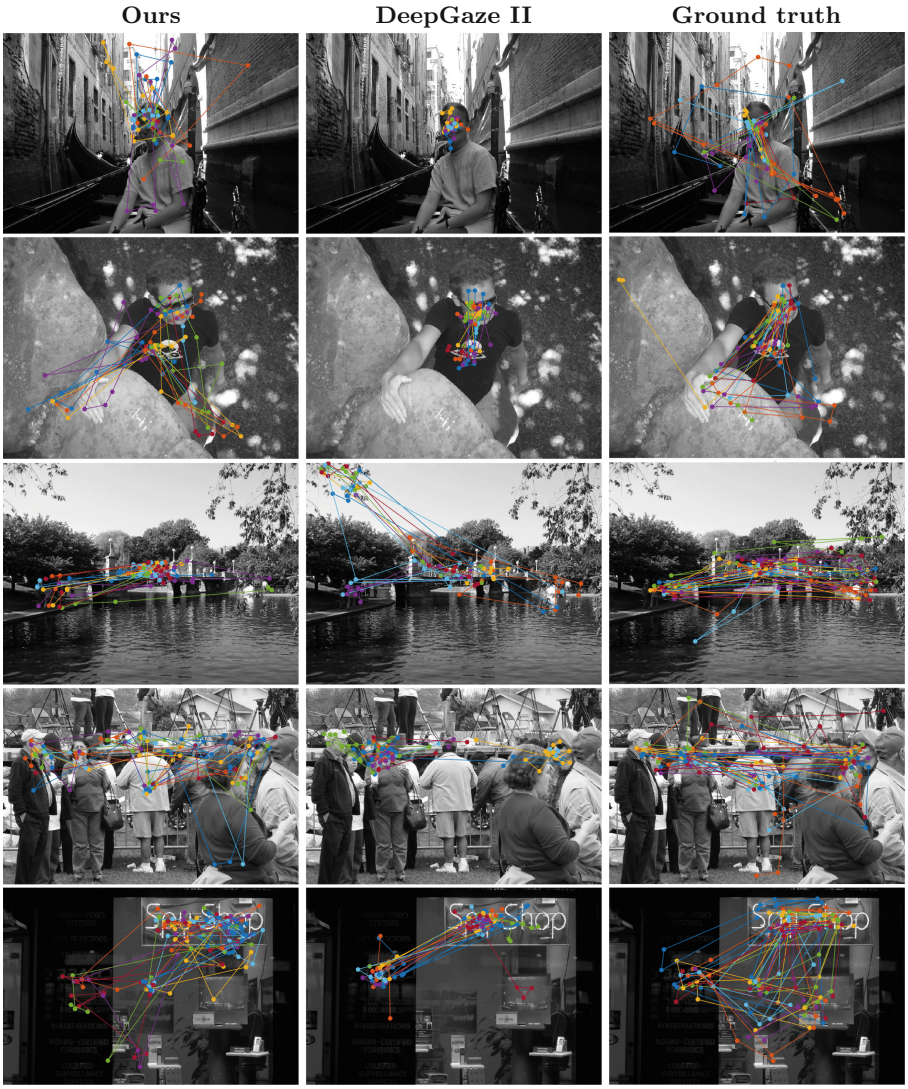


Fig. 6. Examples of scanpaths for the images considered in Fig. 5. Left to right: 15 model-generated scanpaths, from the proposed method, 15 model-generated scanpaths from the DeepGaze II saliency map, 15 scanpaths from actual human fixation sequences (ground-truth). Different colours encode different "observers", either artificial or human.

not faced towards the observer. Finally, as regards text, the last example shows how the whole text region is relevant and not just individual portions of it.

To quantitatively support such insights, the generated scanpaths have been evaluated on each image of the dataset by adopting metrics based on ScanMatch [16] and recurrence quantification analysis (RQA, [2])[2].

ScanMatch is a generalised scanpath comparison method that overcomes the lack of flexibility of the well-known Levenshtein distance (or string edit method) [25]. A similarity score of 1 indicates that the sequences are identical while a score of 0 indicates no similarity. One of the strengths of this method is the ability to take into account spatial, temporal, and sequential similarity between scanpaths; however, as any measure that relies on regions of interest or on a regular grid, it suffers from issues due to quantisation.

Differently, RQA is typically exploited to describe complex dynamical systems. Recently [2] it has been adopted to quantify the similarity of a pair of fixation sequences by relying on a series of measures that are found to be useful for characterizing cross-recurrent patterns [1]. Since we are interested in whether two scanpaths are similar in terms of their fixations sequence, we adopted the determinism and center of recurrence mass (CORM) figures of merit. The determinism provides a measure of the overlap for a sequence of fixations considering the sequential information. The CORM is defined as the distance of the center of gravity of recurrences from the main diagonal in a recurrence plot; small values indicate that the same fixations from the two scanpaths tend to occur close in time.

Results. All the generated scanpaths belonging to our approach and DeepGaze II have been evaluated against human scanpaths for each image. Table 1 reports the average values over all the "observers" related to the same images in the dataset. To quantify the intra-human similarity, an additional measure resulting from the comparison of ground truth scanpaths with themselves is provided.

It must be noted that, in case of DeepGaze II, the adopted model is fine-tuned exactly on the same dataset adopted for testing. Although this clearly introduces bias on the results, it can be seen how the proposed approach outperforms the model without center bias in all three considered metrics. When comparing with the "center bias-aware" model, the ScanMatch result of our approach is worse. In this case, the DeepGaze II output benefits from the addition of a prior distribution estimated over all fixations from the test dataset.

Table 1. Average values (standard deviations) of the considered metrics evaluated over all the artificial and human "observers" related to the same images in the dataset.

	ScanMatch	Determinism	CORM
DeepGazeII w/o CB	0.34 (0.10)	41.16 (16.23)	19.09 (6.21)
DeepGazeII w/ CB	**0.41** (0.07)	50.34 (13.04)	16.39 (4.22)
Ours	0.36 (0.06)	**54.47** (6.54)	**13.75** (2.65)
Ground truth	0.45 (0.05)	59.72 (7.64)	10.02 (2.11)

[2] An implementation is provided at https://github.com/phuselab/RQAscanpath.

4 Conclusive Remarks

Preliminary results show that the "time-aware" scanpaths sampled by taking into account the underlying process of visual attention as unfolding in time considerably differ from those generated by a static attention map; further, they exhibit a dynamics akin to that of scanpaths recorded from human observers.

The model presented here and results so far achieved, albeit simple and preliminary, respectively, bear some consequences. On the one hand, it may suggest a more principled design of visual attention models. A similar perspective has been taken, for instance, in video salience modelling, e.g. [8,15]; nevertheless, static image processing and recognition task could benefit from resorting to dynamics [35]. It is worth noting that the embedding of explicit gaze shift generation is an essential constituent of the model. Too often the design of visual attention models boils down to that of a saliency model. There are of course exceptions to such questionable approach. Le Meur and colleagues [24] have proposed saccadic models as a framework to predict visual scanpaths of observers, where the visual fixations are inferred from bottom-up saliency and oculomotor biases incorporated by gaze shift dynamics are modeled using eye tracking data (cfr. Fig. 1). Yet, there is a limited number of saccadic models available, see [24] for a comprehensive review; generalisation to dynamic scenes have been presented for instance in [6,28]. Also, a "salience free" approach is feasible [40], where steps 2 and 3 can be performed without resorting to an initial salience representation, In [40] generic visual features are exploited via variational techniques under optimality constraints. In this case too a salience map can be obtained *a posteriori* from model-generated fixations [40], but it is just instrumental for comparison purposes [40]. In a similar vein, the maps at the heart of our method do not rely on the concept of saliency as classically conceived. Here, to keep things simple, we have relied on the baseline CLE gaze shift model [3]; yet, one could resort to more complex models, e.g. [4,5].

On the other hand, our approach suggests that fine-grained assessment and benchmarking of models, as surmised in [32], needs to be aware that a static saliency map might not be as predictive of overt attention as it is deemed to be. It is clear that the temporal evolution of the empirical fixation density [7], or its modelling counterpart as proposed here, provides a source of information that is richer than that derived by simply considering its cumulative distribution function at the end of the process.

References

1. Anderson, N.C., Anderson, F., Kingstone, A., Bischof, W.F.: A comparison of scanpath comparison methods. Behav. Res. Methods **47**(4), 1377–1392 (2014). https://doi.org/10.3758/s13428-014-0550-3

2. Anderson, N.C., Bischof, W.F., Laidlaw, K.E.W., Risko, E.F., Kingstone, A.: Recurrence quantification analysis of eye movements. Behav. Res. Methods **45**(3), 842–856 (2013). https://doi.org/10.3758/s13428-012-0299-5
3. Boccignone, G., Ferraro, M.: Modelling gaze shift as a constrained random walk. Phys. A: Stat. Mech. Appl. **331**(1–2), 207–218 (2004)
4. Boccignone, G., Ferraro, M.: Gaze shifts as dynamical random sampling. In: Proceedings of 2nd European Workshop on Visual Information Processing (EUVIP 2010), pp. 29–34. IEEE Press (2010)
5. Boccignone, G., Ferraro, M.: Feed and fly control of visual scanpaths for foveation image processing. Ann. Telecommun. annales des télécommunications **68**(3–4), 201–217 (2013)
6. Boccignone, G., Ferraro, M.: Ecological sampling of gaze shifts. IEEE Trans. Cybern. **44**(2), 266–279 (2014)
7. Boccignone, G., Cuculo, V., D'Amelio, A.: Problems with saliency maps. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019. LNCS, vol. 11752, pp. 35–46. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30645-8_4
8. Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R.: Give ear to my face: modelling multimodal attention to social interactions. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11130, pp. 331–345. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11012-3_27
9. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
10. Bruce, N.D., Wloka, C., Frosst, N., Rahman, S., Tsotsos, J.K.: On computational modeling of visual saliency: examining what's right, and what's left. Vis. Res. **116**, 95–112 (2015)
11. Bylinskii, Z., DeGennaro, E., Rajalingham, R., Ruda, H., Zhang, J., Tsotsos, J.: Towards the quantitative evaluation of visual attention models. Vis. Res. **116**, 258–268 (2015)
12. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2019)
13. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 809–824. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_49
14. Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: experimental data and computer model. J. Vis. **9**(12), 1–15 (2009)
15. Coutrot, A., Guyader, N.: An audiovisual attention model for natural conversation scenes. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 1100–1104. IEEE (2014)
16. Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I.D.: ScanMatch: a novel method for comparing fixation sequences. Behav. Res. Methods **42**(3), 692–700 (2010)
17. Cuculo, V., D'Amelio, A., Lanzarotti, R., Boccignone, G.: Personality gaze patterns unveiled via automatic relevance determination. In: Mazzara, M., Ober, I., Salaün, G. (eds.) STAF 2018. LNCS, vol. 11176, pp. 171–184. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04771-9_14

18. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron (2018). https://github.com/facebookresearch/detectron
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
20. Hu, P., Ramanan, D.: Finding tiny faces. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1522–1530. IEEE (2017)
21. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE (2009)
23. Kummerer, M., Wallis, T.S., Gatys, L.A., Bethge, M.: Understanding low-and high-level contributions to fixation prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4789–4798 (2017)
24. Le Meur, O., Coutrot, A.: Introducing context-dependent and spatially-variant viewing biases in saccadic models. Vis. Res. **121**, 72–84 (2016)
25. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**, 707–710 (1966)
26. Li, X., Wang, W., Hou, W., Liu, R.Z., Lu, T., Yang, J.: Shape robust text detection with progressive scale expansion network. arXiv preprint arXiv:1806.02559 (2018)
27. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
28. Napoletano, P., Boccignone, G., Tisato, F.: Attentive monitoring of multiple video streams driven by a Bayesian foraging strategy. IEEE Trans. Image Process. **24**(11), 3266–3281 (2015)
29. Nguyen, T.V., Zhao, Q., Yan, S.: Attentive systems: a survey. Int. J. Comput. Vis. **126**(1), 86–110 (2018)
30. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. Prog. Brain Res. **155**, 23–36 (2006)
31. Rothkegel, L.O., Trukenbrod, H.A., Schütt, H.H., Wichmann, F.A., Engbert, R.: Temporal evolution of the central fixation bias in scene viewing. J. Vis. **17**(13), 3 (2017)
32. Schütt, H.H., Rothkegel, L.O., Trukenbrod, H.A., Engbert, R., Wichmann, F.A.: Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. J. Vis. **19**(3), 1 (2019)
33. Tatler, B.W., Hayhoe, M.M., Land, M.F., Ballard, D.H.: Eye guidance in natural scenes: reinterpreting salience. J. Vis. **11**(5), 1–23 (2011)
34. Tatler, B., Vincent, B.: The prominence of behavioural biases in eye guidance. Vis. Cogn. **17**(6–7), 1029–1054 (2009)
35. Tavakoli, H.R., Borji, A., Anwer, R.M., Rahtu, E., Kannala, J.: Bottom-up attention guidance for recurrent image recognition. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3004–3008. IEEE (2018)
36. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. **113**(4), 766 (2006)

37. Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. J. Vis. **9**(7), 4 (2009)
38. Xia, C., Han, J., Qi, F., Shi, G.: Predicting human saccadic scanpaths based on iterative representation learning. IEEE Trans. Image Process. **28**(7), 3502–3515 (2019)
39. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
40. Zanca, D., Gori, M.: Variational laws of visual attention for dynamic scenes. In: Advances in Neural Information Processing Systems, pp. 3823–3832 (2017)
41. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
42. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)