Problems with saliency maps

Giuseppe Boccignone[0000−0002−5572−0924], Vittorio Cuculo[0000−0002−8479−9950], and Alessandro D'Amelio[0000−0002−8210−4457]

PHuSe Lab - Dipartimento di Informatica, University of Milan
via Celoria 18, Milano, 20133 Italy
{giuseppe.boccignone,vittorio.cuculo,alessandro.damelio}@unimi.it
phuselab.di.unimi.it

Abstract. Despite the popularity that saliency models have gained in the computer vision community, they are most often conceived, exploited and benchmarked without taking heed of a number of problems and subtle issues they bring about. When saliency maps are used as proxies for the likelihood of fixating a location in a viewed scene, one such issue is the temporal dimension of visual attention deployment. Through a simple simulation it is shown how neglecting this dimension leads to results that at best cast shadows on the predictive performance of a model and its assessment via benchmarking procedures.

Keywords: Saliency model · Visual attention · Gaze deployment.

1 Introduction

Many efforts have been devoted in the past decade to the computational modelling of visual salience [6,7,8,10,9,38], and recently large breakthroughs have been achieved on benchmarks by resorting to deep neural network models [10].

Saliency models are appealing since, apparently, they represent a straightforward operational definition of visual attention - the allocation of visual resources to the viewed scene [8]: they take an image $\mathbf{I}(\mathbf{r})$ as input, and return topographic maps $\mathcal{S}(\mathbf{r})$ indicating the salience at each location $\mathbf{r} = (x, y)$ in the image, namely the likelihood of fixating at $\mathbf{r}$. 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 [29].

Yet, salience modelling and benchmarking are most often handled in an elusive way, which casts doubts on a straightforward interpretation of results so far achieved [33,8,24,22,30]. Beyond the long debated controversy concerning the bottom-up vs. top-down nature of eye guidance control [17,33], factors such as context [36], spatial biases [34], affect and personality [16], dynamics of attention deployment [31,30] are likely to play a key role and might contribute in subtle ways to effectiveness and performance of saliency models [33,24,22,30]. Some controversial aspects related to salience definition and modelling are discussed in Section 2. In particular, the temporal unfolding of factors [30] involved in salience making has been overlooked, with some exceptions in video processing (e.g. [5,13]) but largely neglected in static images.

The hitherto underestimated point we are making in this note is that by explicitly taking into account temporal unfolding provides useful conceptual insights on the actual predictive capability of saliency models, with practical consequences on their use and benchmarking.

By and large, 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). In a different vein, here we operationally take into account temporal aspects of attention deployment captured by a time-varying fixation map (Section 3). Through a simple experiment, we show (Section 4) that in such way the actual sampling of gaze shifts, namely *how* we actually allocate visual resources onto the scene (i.e., the scanpath), can depart from that achieved by classic analyses.

2 The salience conundrum: background and motivation

Saliency models. The notion of salience originates in visual attention research (e.g., [21]). In the case of overt visual attention, actual eye movements are involved. Eye movements obviously occur according to a continuous dynamics but their spatial and velocity characteristics allow to classify them as fixations, saccades and smooth pursuit of moving objects. Fixations and pursuit aim to bring or keep objects of interest onto the fovea where the visual acuity is maximum, whilst saccades are ballistic shifts in eye position, allowing to jump from one location of the viewed scene to another. When considering overt attention involving gaze, then the aim of a computational model of attentive eye guidance is to answer the question *Where to Look Next?* by providing an account of the mapping from visual data of a natural scene, say $\mathbf{I}$ (the raw data representing either a static picture or a stream of images), to a sequence of time-stamped gaze locations $(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \cdots$, namely $\mathbf{I} \mapsto \{\mathbf{r}_{F_1}, t_1; \mathbf{r}_{F_2}, t_2; \cdots\}$. The common practice to derive such mapping is to conceive it as a two stage procedure: (i) Compute a suitable perceptual representation $\mathcal{W}$, i.e., $\mathbf{I} \mapsto \mathcal{W}$; (ii) Use $\mathcal{W}$ to generate the scanpath, $\mathcal{W} \mapsto \{\widetilde{\mathbf{r}}_F(1), \widetilde{\mathbf{r}}_F(2), \cdots\}$ (where we have adopted the compact notation $(\widetilde{\mathbf{r}}_{F_n}, t_n) = \widetilde{\mathbf{r}}_F(n)$).

Stimulus salience is one such perceptual representation $\mathcal{W}$. It is the driving force behind bottom-up or "exogenous" attention control, driven by low-level scene properties (brightness, colour, oriented edges, motion contrasts [19]) and independently of the internal mental state of the observer. Indeed, for the most part, the first computable models for the prediction of eye fixation locations in images relied on a "saliency map", $\mathcal{S}$ a topographic representation indicating *where* one is likely to look within the viewed scene [19], 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 $\mathbf{I}$.

By overviewing the field [33,6,7,8], it is easily recognised that computational modelling of visual attention has been mainly concerned with stage (i), that is calculating $\mathcal{W} = \mathcal{S}$. As to stage (ii), namely $\mathcal{S} \mapsto \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$, which actually brings in the question of *how* we look rather than *where*, it is seldom taken into account.

Model performance. An issue that straightforwardly raises is how to measure and benchmark the performance of a saliency model accounting for the map $\mathbf{I} \mapsto \mathcal{S}$. The general idea is to measure the capability of the model output, namely the saliency map $\mathcal{S}$, to predict fixations (notice: *as if* they were performed). To such end, eye fixations $\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}$ are typically used as to derive the ground-truth. These are collected in an eye-tracking experiment involving $s = 1 \cdots N_S$ subjects on a chosen data set $\{\mathbf{I}^i\}$ of $i = 1 \cdots N_I$ images (or videos). Some metrics use the original binary location map of fixations, say $\mathcal{M}^B$. Alternatively, the discrete fixations can be converted into a continuous distribution, a fixation map (a.k.a *heat map* or *attention map* when fixations are weighted by fixation time), $\mathcal{M}^D$ [9]. Precisely, for each stimulus $\mathbf{I}^i$ the map

$$\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}_{s=1}^{N_S} \mapsto \mathcal{M}^{D(i)}, \tag{1}$$

is computed as an empirical fixation density (e.g.,[23,25]); see Fig. 1 below. Eventually, a metric is evaluated either in the form $\mu(\mathcal{S}, \mathcal{M}^B)$ or $\mu(\mathcal{S}, \mathcal{M}^D)$, the result being a number assessing the similarity or dissimilarity between $\mathcal{S}$, and $\mathcal{M}$ (for an in-depth presentation, see Bylinskii *et al.*[9]).

The many facets of salience (and benchmarking) Despite the considerable successes that salience has had in predicting fixations at above-chance levels, it has become increasingly clear that prediction requires high-level, semantically-meaningful elements (e.g. faces, objects and text [11,12]). Thus, prominent models of attention control posit a dichotomy between bottom-up and top-down, "endogenous" control, the latter being determined by current selection goals; in this case spotted items are selected in terms of their goal relevance, rather than physical salience. In the visual attention realm when top-down (relevance) and bottom-up (saliency) mechanisms are combined for eye guidance, the resulting map is termed priority map [17].

In a different vein, 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 saliency models [6,18,7,8,10,9]. 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]. Indeed, the success of deep networks exploiting convolutional filters that have been learned on other tasks, for instance object recognition in the ImageNet dataset, provides practical evidence of the usefulness of high-level image features for prediction purposes [10,24]. In recent evaluations on what should be the next steps in salience modelling and assessment [8,10], it has been shown that a large improvement in predictive performance could be gained by specifically addressing semantic issues such as actions in a scene, relative importance to different faces, informativeness of text, targets of gaze.

Meanwhile, such practice somehow limits a straightforward interpretation of benchmarking results so far achieved; thus, disentangling the different levels of control to understand to what extent fixations in free viewing are driven by low-level features or by high-level features is recently growing up as a research line *per se* [24,22,30]

3 Temporal unfolding of fixation allocation

Crucially, saliency maps do not account for temporal dynamics. They are by and large spatially evaluated across all fixations, precisely by comparing to maps $\mathcal{M}^B$, or $\mathcal{M}^D$ derived from fixations accumulated in time after the stimulus onset until the end of the trial (Eq. 1).

As a matter of fact, surmising that $\mathcal{S}$ is predictive of human fixations does not entail an actual mechanism of fixation generation, $\mathcal{S}_i \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(1), \widetilde{\mathbf{r}}_F^{(s,i)}(2), \cdots\}$ to be compared against actual fixation sequences $\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}$. The assessment of the predictive capability of a model is just to be understood as the indirect measurement of any metric μ as introduced above. When using the mapping of Eq. 1, it is implicitly assumed that fixations, once collected, are exchangeable with respect to time ordering $\{1, \cdots, n\}$, namely

$$\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots \mathbf{r}_F^{(s,i)}(n)\} = \{\mathbf{r}_F^{(s,i)}(\pi(1)), \mathbf{r}_F^{(s,i)}(\pi(2)), \cdots, \mathbf{r}_F^{(s,i)}(\pi(n))\}, \tag{2}$$

$\forall \pi \in \Pi(n)$ where $\Pi(n)$ is the group of permutations of $\{1, \cdots, n\}$. This assumption implies that any dynamical law $\widetilde{\mathbf{r}}_F^{(s,i)}(t) = f(\widetilde{\mathbf{r}}_F^{(s,i)}(t-1), \mathcal{W}_i)$ that takes as input the perceptual representation of the i-th image and the previous fixation location (as a system state) and returns the next location of fixation as its output is dismissed. However, dynamics is important in many respects. For instance, there is evidence for the existence of systematic tendencies in oculomotor control [34]: eyes are not equally likely to move in any direction. Yet, apart from the well known center bias [32], motor biases can be actually taken into account only when scanpath generation is performed.

In such perspective, Le Meur and colleagues [26] have proposed saccadic models as a new framework to predict visual scanpaths of observers while they freely watch static images. In such models the visual fixations are inferred from bottom-up saliency and oculomotor biases (captured as saccade amplitudes and saccade orientations) that are modeled using eye tracking data. Performance of these models can be evaluated either by directly comparing the generated scanpaths to human scanpaths or by computing new saliency maps, in the shape of densities from model generated fixations. There is a limited number of saccadic models available, see [26] for a comprehensive review; generalisation to dynamic scenes have been presented for instance in [4,28]. A remarkable result obtained by saccadic models is that by using simulated fixations $\{\widetilde{\mathbf{r}}_F^{(s,i)}(1), \widetilde{\mathbf{r}}_F^{(s,i)}(2), \cdots\}$ to generate a model-based fixation map, the latter has higher predictive performance than the raw salience map $\mathcal{S}$, in terms of similarity/dissimilarity μ with respect to human fixation maps. Beyond the improvement, it is worth noting that even in this case the model-generated attention map is eventually obtained *a posteriori*, as a 2-D spatial map of accumulated fixations. Such problem is somehow attenuated when dynamic stimuli (videos) are taken into account, though, the temporal unfolding as learned in a data-driven way presents complex albeit structured temporal patterns [5,14], that deserve being taken into consideration.

In a different vein, recent work by Schutt *et al.* [30] has for the first time considered the temporal evolution of the fixation density in the free viewing of static scenes. They provide evidence for a fixation dynamics which unfolds into three phases:

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.

Beyond the theoretical insights offered by their analyses, by monitoring the performance of the empirical fixation density over time, they also pave the way to a more subtle and principled approach to unveil the actual predictive performance of saliency models [30].

Based on their approach, we propose a complementary analysis that relies on model-generated scanpaths, i.e. actual prediction. More precisely, we ask the following: do model-generated scanpaths differ from human scanpaths in the free viewing of static scenes when 1) the scanpath is generated by taking into account the three phases described above as opposed to when 2) the scanpath is generated by only taking into account the final fixation density?

In the work presented here, we use the time-varying fixation density as the attention map that moment-to-moment feeds the gaze shift dynamics. The main motivation is in the very fact that we want to assess differences rising at the oculomotor behavior while being free from any saliency model specific assumption. In brief we do the following:

Step 1 Compute three different empirical fixation density maps $\mathcal{M}_k^{D(i)}$ accounting for phases $k = 1, 2, 3$ above, by aggregating all the human fixations performed in the corresponding time window:

$$\{\mathbf{r}_F^{(s,i)}(m_{k-1}+1), \cdots, \mathbf{r}_F^{(s,i)}(m_k)\}_{s=1}^{N_S} \mapsto \mathcal{M}_k^{D(i)}, \quad k = 1, 2, 3. \tag{3}$$

Step 2. Generate "subject" fixations depending on the three-phase unfolding defined above, by relying on a saccadic model $\mathbf{r}_F^{(s,i)}(n) = f(\mathbf{r}_F^{(s,i)}(n-1), \mathcal{W}(k)_i)$:

$$\mathcal{M}_k^{D(i)} \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(m_{k-1}+1), \cdots, \widetilde{\mathbf{r}}_F^{(s,i)}(m_k)\} = \mathcal{R}t_k^{(s,i)}, \;\; k = 1, 2, 3 \tag{4}$$

with $\mathcal{W}(k)_i = \mathcal{M}_k^{D(i)}$ being the phase-dependent perceptual representation of image i, so to obtain 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 comparison purposes, in the same way, but only by relying on the overall final fixation map $\mathcal{M}^{D(i)}$, we perform the mapping $\mathcal{M}^{D(i)} \mapsto \mathcal{R}s^{(s,i)}$, which represents the typical output of a saccadic model.

4 Simulation

Dataset The adopted dataset is a publicly available one [20], that consists of eye tracking data (240Hz) recorded from $N_S = 15$ viewers during a free-viewing experiment involving 1003 natural images. The stimuli were presented at full resolution for 3 seconds. The raw eye tracking data were classified in fixations and saccades by adopting an acceleration threshold algorithm [20].

Evaluation As described in the Method section, we generated four different attention maps for each image $\mathbf{I}^i$ of the dataset. Three of these are the temporal density fixation maps $\mathcal{M}_1^{D(i)}, \mathcal{M}_2^{D(i)}, \mathcal{M}_3^{D(i)}$, with $t_{m_1} = 1$, $t_{m_2} = 2$ and $t_{m_3} = 3$ seconds (Eq. 3); the fourth is the classic, cumulative $\mathcal{M}^{D(i)}$ map. Fig. 1 shows one example. These were used to support the generation of $N_S = 15$ scanpaths for both the temporal (Eq. 4) and the classic approach, collected into the sets $\mathcal{R}t^{(i)}$ and $\mathcal{R}s^{(i)}$, respectively. To such end we exploit the Constrained Levy Exploration (CLE [3])[1] saccadic model that has been widely used for evaluation purposes, e.g.,[26,37]. Briefly, the CLE considers the gaze motion as given by the stochastic dynamics of a Lévy forager moving under the influence of an external force (which, in turn, depends on a salience or attention potential field). Namely, at time t the transition from the current position $\mathbf{r}(t)$ to a new position $\mathbf{r}_{new}(t)$, $\mathbf{r}(t) \to \mathbf{r}_{new}(t)$, is given by

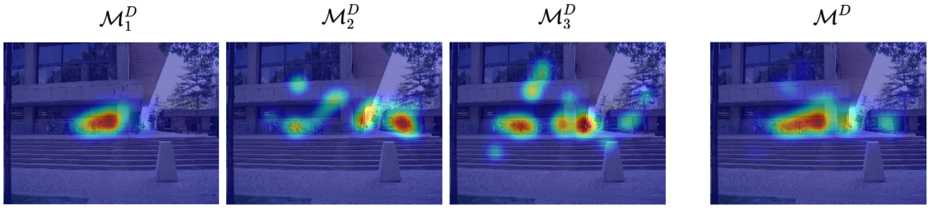

Fig. 1. Example of different fixation density maps for a specific image. From left to right: the three temporal distribution maps obtained from fixations collected at seconds 1, 2 and 3, respectively, overlapped on the original stimulus; the standard fixation map resulting from the aggregation of all fixations available at the end of the eye-tracking procedure. The latter map is the one typically exploited in saliency modelling and benchmarking.

$$\mathbf{r}_{new}(t) = \mathbf{r}(t) + \mathbf{g}(\mathcal{W}(\mathbf{r}(t))) + \boldsymbol{\eta}. \tag{5}$$

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

The trajectory of the variable $\mathbf{r}$ is determined by a deterministic part $\mathbf{g}$, the drift - relying upon salience or fixation density -, and a stochastic part $\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a random vector sampled from a heavy-tailed distribution, accounting for motor biases (cfr., the Appendix for a quick recap and [3] for theoretical details).

Figure 2 shows CLE generated scanpaths, compared against the actual set of human scanpaths $\mathcal{R}^{(i)} = \{\mathbf{r}_F^{(i)}(1), \cdots, \mathbf{r}_F^{(i)}(m_3)\}$. The example shows at a glance that when attention deployment is unfolded in time, the predicted scanpaths more faithfully capture the dynamics of actual scanpaths than the dynamics of those generated via the "freezed" map. To quantitatively support such insight, the quality of $\mathcal{R}t^{(i)}$ and $\mathcal{R}s^{(i)}$ has been evaluated on each image i of the dataset by adopting metrics based on the ScanMatch [15] and the recurrence quantification analysis (RQA, [2])[2].

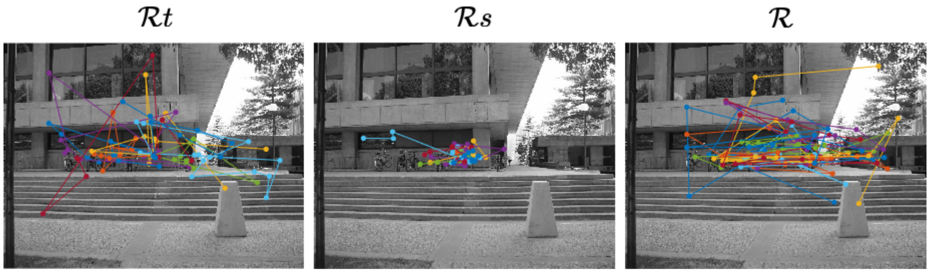

Fig. 2. Scanpaths for the image in Fig. 1. Left to right: 15 model-generated scanpaths, via Eq. 5 from the temporally unfolded fixation maps, 15 model-generated scanpaths from the standard fixation map, 15 scanpaths from actual human fixation sequences (ground-truth). Different colours encode different "observers"(artificial or human.

ScanMatch is a generalised scanpath comparison method that overcomes the lack of flexibility of the well-known Levenshtein distance (or string edit method) [27]. It consists of a preliminary step, where two scanpaths are spatially and temporally binned and then re-coded to create sequences of letters that preserve fixation location, time and order. These are then compared adopting the Needleman-Wunsch sequence alignment algorithm, widely used to compare DNA sequences. The similarity score is given by the optimal route throughout a matrix that provides the score for all letter pair substitutions and a penalty gap. A similarity score of 1 indicates that the sequences are identical; 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 grid, it suffers from quantisation issues.

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

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

relying on a series of measures that are found to be useful for characterizing cross-recurrent patterns [1]. RQA calculates the cross-recurrence for each fixation of two scanpaths, resulting in the construction of the so-called recurrence plot: two fixations are cross-recurrent if they are close together in terms of their Euclidean distance. 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) measures. The determinism represents the percentage of cross-recurrent points that form diagonal lines in a recurrence plot; it 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 both scanpaths tend to occur close in time.

Results All the generated scanpaths belonging to $\mathcal{R}t$ and $\mathcal{R}s$ have been evaluated against the human ones $\mathcal{R}$ 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 $\mathcal{R}$ with itself is provided. It can be noticed that the temporal approach outperforms the static one in all the three adopted metrics. Remarkably, as regards the determinism, the percentage of overlapping sequences when adopting the temporal approach is higher than that resulting from the comparison among human scanpaths. This would suggest that a high inter-subject variability occurs when looking at the same stimulus, and that the adoption of temporal maps does extract common behaviour among the observers, resulting in a lower spread of fixation locations.

	ScanMatch	**Determinism**	**CORM**
$\mathcal{R}s$ **vs.** $\mathcal{R}$	0.39 (0.08)	58.08 (11.18)	19.95 (5.90)
$\mathcal{R}t$ **vs.** $\mathcal{R}$	**0.43** (0.05)	**61.65** (8.51)	**15.26** (3.58)
$\mathcal{R}$ **vs.** $\mathcal{R}$	0.49 (0.05)	59.61 (7.71)	10.0 (2.09)

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.

5 Conclusive remarks

In this note by resorting to a straightforward simulation of scanpath generation, evidence has been given that: (i) the scanpaths sampled by taking into account the underlying process of visual attention unfolding in time (dynamic attention map) considerably differ from those generated by a static attention map; (ii) "time-aware" model-based scanpaths exhibit a dynamics akin to that of scanpaths recorded from human observers.

It should be intuitively apparent 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 $\int_{t_0}^{T} \mathcal{M}_t^{D(i)} dt$. Yet, this very fact is by and large neglected in the saliency modelling practice. It has to be said that this pitfall is somehow mitigated when dynamic stimuli (videos) are taken into account. Though, a large body of research is still flourishing in pursuit of adequate computational models of salience in static images.

The analysis reported here bear some consequences.

On the one hand, it may suggest a more principled design of visual attention models specially when time dimension is crucial for the analysis. Here to keep the discussion simple, we have straightforwardly used empirical fixation density maps $\mathcal{M}_t^{D(i)}$ derived via the mapping (3). However, nothing prevents from building models based on a chain of sub-models, each contributing to the final scanpath, thus following the same route we have outlined above. For example, the three-stage processing suggested in [30], could be accounted for by (1) a center-bias model, (2) a context/layout model, and (3) an object-based model, respectively. A similar perspective has been taken, for instance, in video salience modelling; nevertheless, static image processing and recognition task could benefit from resorting to dynamics [35].

On the other hand, the approach could be used for fine-grained assessment of models as surmised in [30]; hence, being aware that a static saliency map might not be as predictive of overt attention as it is deemed to be.

Appendix: The Lévy forager

The Lévy forager's dynamics formalised in Eq.5 can be written

$$\mathbf{r}_{new}(t) = \mathbf{r}(t) - \nabla V + \boldsymbol{\eta}, \tag{6}$$

so that the new gaze position is determined by: a) the gradient of V, the external force field shaped by the perceptual landscape, $V(\cdot, t)$ being defined as the time varying scalar field

$$V(x, y, t) = \exp(-\tau_V \mathcal{W}(x, y, t)), \tag{7}$$

b) the stochastic vector $\boldsymbol{\eta}$ with components

$$\eta_x = l \cos(\theta), \qquad \eta_y = l \sin(\theta), \tag{8}$$

where the angle θ represents the flight direction and l is the jump length. Direction and length are sampled from the uniform and α-stable distribution, respectively:

$$\theta \sim Unif(0, 2\pi), \tag{9}$$

$$l \sim \varphi(\mathcal{W}) f(l; \alpha, \beta, \gamma, \delta). \tag{10}$$

Along the extensive stage, θ and l summarise the internal action choice of the forager and the function $\varphi(\mathcal{W})$ modifies the pure Levy flight, since the probability to move from one site to the next site depends on the "strength" of a bond

$$\varphi(\mathcal{W}) = \frac{\exp(-\beta_P(\mathcal{W}(\mathbf{r}(t)) - \mathcal{W}(\mathbf{r}_{new}(t))))}{\sum_{\mathbf{r}'_{new}} \exp(-\beta_P(s(\mathbf{r}(t)) - \mathcal{W}(\mathbf{r}'_{new}(t))))} \tag{11}$$

that exists between them. The shift proposal is weighed up according to an accept/reject Metropolis rule that depends on the perceptual gain $\varDelta\mathcal{W}$ and on "temperature" T [3]. The values of T determine the amount of randomness in scanpath generation. If no suitable shift $\mathbf{r}(t)_{new}$ has been selected, the current fixation point $\mathbf{r}(t)$ is retained.

References

1. Anderson, N.C., Anderson, F., Kingstone, A., Bischof, W.F.: A comparison of scanpath comparison methods. Behavior research methods **47**(4), 1377–1392 (2015)
2. Anderson, N.C., Bischof, W.F., Laidlaw, K.E., Risko, E.F., Kingstone, A.: Recurrence quantification analysis of eye movements. Behavior research methods **45**(3), 842–856 (2013)
3. Boccignone, G., Ferraro, M.: Modelling gaze shift as a constrained random walk. Physica A: Statistical Mechanics and its Applications **331**(1-2), 207–218 (2004)
4. Boccignone, G., Ferraro, M.: Ecological sampling of gaze shifts. IEEE Trans. on Cybernetics **44**(2), 266–279 (Feb 2014)
5. 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.) Computer Vision – ECCV 2018 Workshops, pp. 331–345. Springer International Publishing, Cham (2019)
6. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 185–207 (2013)
7. 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. Vision research **116**, 95–112 (2015)
8. Bylinskii, Z., DeGennaro, E., Rajalingham, R., Ruda, H., Zhang, J., Tsotsos, J.: Towards the quantitative evaluation of visual attention models. Vision research **116**, 258–268 (2015)
9. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. on Pattern Analysis and Machine Intelligence **41**(3), 740–757 (March 2019)
10. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: European Conference on Computer Vision. pp. 809–824. Springer (2016)
11. Cerf, M., Frady, E., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision **9**(12) (2009)
12. Clavelli, A., Karatzas, D., Lladós, J., Ferraro, M., Boccignone, G.: Modelling task-dependent eye guidance to objects in pictures. Cognitive Computation **6**(3), 558–584 (2014)
13. Coutrot, A., Guyader, N.: An efficient audiovisual saliency model to predict eye positions when looking at conversations. In: 23rd European Signal Processing Conference. pp. 1531–1535 (Aug 2015)

14. Coutrot, A., Guyader, N.: How saliency, faces, and sound influence gaze in dynamic social scenes. Journal of vision **14**(8), 5–5 (2014)
15. Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I.D.: Scanmatch: A novel method for comparing fixation sequences. Behavior Research Methods **42**(3), 692–700 (2010)
16. Cuculo, V., D'Amelio, A., Lanzarotti, R., Boccignone, G.: Personality gaze patterns unveiled via automatic relevance determination. In: Federation of International Conferences on Software Technologies: Applications and Foundations. pp. 171–184. Springer (2018)
17. Egeth, H.E., Yantis, S.: Visual attention: Control, representation, and time course. Annual review of psychology **48**(1), 269–297 (1997)
18. Furnari, A., Farinella, G.M., Battiato, S.: An experimental analysis of saliency detection with respect to three saliency levels. In: European Conference on Computer Vision. pp. 806–821. Springer (2014)
19. 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**, 1254–1259 (1998)
20. 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)
21. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol **4**(4), 219–27 (1985)
22. Kong, P., Mancas, M., Thuon, N., Kheang, S., Gosselin, B.: Do deep-learning saliency models really model saliency? In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 2331–2335. IEEE (2018)
23. Kümmerer, M., Wallis, T.S., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences **112**(52), 16054–16059 (2015)
24. 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)
25. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behavior Research Methods **45**(1), 251–266 (Mar 2013)
26. Le Meur, O., Coutrot, A.: Introducing context-dependent and spatially-variant viewing biases in saccadic models. Vision Research **121**, 72–84 (2016)
27. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)
28. Napoletano, P., Boccignone, G., Tisato, F.: Attentive monitoring of multiple video streams driven by a bayesian foraging strategy. IEEE Trans. on Image Processing **24**(11), 3266 – 3281 (Nov 2015)
29. Nguyen, T.V., Zhao, Q., Yan, S.: Attentive systems: A survey. International Journal of Computer Vision **126**(1), 86–110 (2018)
30. 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. Journal of vision **19**(3), 1–1 (2019)
31. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: Effects of scale and time. Vision research **45**(5), 643–659 (2005)
32. Tatler, B.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision **7**(14) (2007)

33. Tatler, B., Hayhoe, M., Land, M., Ballard, D.: Eye guidance in natural vision: Reinterpreting salience. Journal of vision **11**(5) (2011)
34. Tatler, B., Vincent, B.: The prominence of behavioural biases in eye guidance. Visual Cognition **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. Psychological review **113**(4), 766 (2006)
37. Xia, C., Han, J., Qi, F., Shi, G.: Predicting human saccadic scanpaths based on iterative representation learning. IEEE Transactions on Image Processing pp. 1–1 (2019)
38. Zhang, J., Malmberg, F., Sclaroff, S.: Visual Saliency: From Pixel-Level to Object-Level Analysis. Springer (2019)