SCIENTIFIC REPORTS
nature research

OPEN

Gravitational models explain shifts on human visual attention

Dario Zanca[1]✉, Marco Gori[2,3], Stefano Melacci[2] & Alessandra Rufa[1]

Visual attention refers to the human brain's ability to select relevant sensory information for preferential processing, improving performance in visual and cognitive tasks. It proceeds in two phases. One in which visual feature maps are acquired and processed in parallel. Another where the information from these maps is merged in order to select a single location to be attended for further and more complex computations and reasoning. Its computational description is challenging, especially if the temporal dynamics of the process are taken into account. Numerous methods to estimate saliency have been proposed in the last 3 decades. They achieve almost perfect performance in estimating saliency at the pixel level, but the way they generate shifts in visual attention fully depends on winner-take-all (WTA) circuitry. WTA is implemented by the biological hardware in order to select a location with maximum saliency, towards which to direct overt attention. In this paper we propose a gravitational model to describe the attentional shifts. Every single feature acts as an attractor and the shifts are the result of the joint effects of the attractors. In the current framework, the assumption of a single, centralized saliency map is no longer necessary, though still plausible. Quantitative results on two large image datasets show that this model predicts shifts more accurately than winner-take-all.

Despite the huge amount of data that reaches the human eye every second[1], neuronal hardware is insufficient to process it all at once. Indeed vision is. This crucial set of information is collected and forwarded to intermediate and higher levels of processing. The study of the visual attention mechanism has been in the spotlight for the past 3 decades[2]. It is at the crossroads of different disciplines such as psychology[3,4], cognitive neuroscience[5,6], computer vision[7–12]. Although great advances have been produced, we are still far from defining a model that approximates human capabilities. Models of human visual attention are of great interest for the scientific community. They help researchers to understand the cognitive mechanisms of visual selection, which happens to be very intertwined with top-down processes[13–16], but at the same time they provide a wide range of applications such as in marketing[17,18], video compression[19] or virtual reality[20]—just to name a few. What makes modeling of human visual attention challenging is its inherently dynamic nature. Subsequent shifts in human attention are highly correlated with previous overt gaze shifts, as well as with the dynamics with which the scene itself changes[21], the neural correlate of which, has been demonstrated to reside in a common population of neurons lying in frontal eye field (FEF)[22,23].

Current approaches in modeling human visual attention are based on the so-called saliency hypothesis[24]. It postulates the existence of a saliency map whose function is to guide attention and gaze towards the most conspicuous regions of the visual scene. This hypothesis has received numerous independent experimental confirmations[25]. Following the approach traced by the seminal works of Itti and Koch[7], current approaches concentrate their efforts on the problem of learning saliency from human data. Attention models generally yield an output saliency map that indicates the probability of fixating each location of the space[12,26]. However, this does not model the temporal dynamics of the process in terms of the temporal order of fixations, i.e. scanpaths. For this reason, authors often assume that a circuitry[7,24] of winner-take-all (WTA) is implemented by the human biological hardware to generate sequences of fixations, starting from saliency maps. This approach, however, still fails to provide a continuous dynamic of the process. Scanpath simulations are poor, not plausible and, consequently, not reliable for applications.

In this paper, we propose and validate a different model for generating attentional shifts over time. Our description is minimal both in the perceptual scheme and in the mathematical formulation. In the proposed framework, the encoding of the oculomotor command needs a simple neuronal hardware, with the very mild

[1]Department of Medicine, Surgery and Neuroscience, University of Siena, 53100 Siena, Italy. [2]Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy. [3]Inria, CNRS, I3S, Université Côte d'Azur, Maasai, Côte d'Azur, France. ✉email: dario.zanca@unisi.it

assumption of units that perform sum operation, without the need of backward flows, leading to a plausible and straightforward biological implementation. The literature indicates that the V1 area is important for the conformation of the bottom-up saliency[27,28], while other associative areas such as V4, FEF and supplementary eye field (SEF)[29] receive signals simultaneously from both the V1 area and deeper layers[30]. Neurons in the V1 area can encode principal and independent components extracted directly from the visual input[28], and the response magnitude of such neurons is greater when the stimulus is distinct from its surrounding[27,28]. Moreover, some studies confirm the effect of the V1 area on bottom-up visual attention in free-viewing scenes[25]. Following the outcome of scientific studies on the V1 area of the brain and its relationships in bottom-up visual attention, we make the choice of minimal and naturalistic design, taking into account only basic features to represent the input signal, such as color, intensity and orientation gradients.

The main function of visual systems is to capture as much information as possible given limited resources. Its main sensory limitation is spatial resolution which varies across the retina depending on the arrangement of photoreceptor mosaic and their receptive fields. Under this considerations, we provide a mathematical formulation in which conspicuous features are modeled as masses that compete to attract visual attention. We derive laws that regulate attentional shifts through a gravitational model. The resulting shifts will be a consequence of gravitational attractions, together with a mechanism of inhibition of return (IOR), part of the visual foraging behavior, that allows the model to explore the whole scene. While, in principle, such a description is formulated independently from biological mechanisms, the biological plausibility of the proposed model and a sketch of the neuronal hardware needed for realizing the underlying computation is given.

The output of the model is a continuous function that describes the trajectory of the focus of attention. It is worth noting that the proposed framework makes it easy to introduce additional visual features. External signals can be introduced to model the field of attention-grabbing masses to align it with specific tasks. Our approach relies on differential laws of motion, and it naturally provides the temporal dynamic of the exploration of the scene. Measures of scanpath similarity are adopted to measure plausibility of the model and closeness to real human data. In particular, three different metrics in literature have been shown to be robust: string-edit distance (SED)[31–34], time-delay embeddings (TDE)[35] and scaled time delay-embeddings (STDE)[34,36].

Results

A gravitational model of visual attention. A generic stream of visual input is defined on the domain $\mathbb{D} = \mathbb{R} \times \mathbb{T}$, where the subset $\mathbb{R} \subset \mathbb{R}^2$ represents the retina coordinates while $\mathbb{T} \subset \mathbb{R}$ is the temporal domain. The visual attention scanpath is the trajectory $a(t) : \mathbb{T} \to \mathbb{R}$, being $t \in \mathbb{T}$ the time index. Attention is assumed to be driven by the attraction triggered by a collection of N relevant visual features. Let $f_i : \mathbb{D} \to \mathbb{R}$ be the function associated with the activation of a visual feature i modeling the presence of a certain property in a pixel of the input stream, where $i \in \{1, \ldots, N\}$. Larger values of $f_i(x, t)$ correspond with more evident presence of the visual feature in $(x, t) \in \mathbb{D}$, being x the pixel coordinates. Let us assume to have the use of a number of f_i's, each of them associated to different properties of the input stream.

Inspired by the behaviour of gravitation fields, that naturally embed the idea of attraction, we model the visual attention scanpath as the motion of a unitary mass subject to the gravitational attraction of a distribution of masses μ, associated to the visual features, $\mu : \mathbb{D} \to \mathbb{R}$. In particular, we define $\mu(x, t) = \sum_i \mu_i(x, t)$, being μ_i the mass associated to feature f_i, that is

$$\mu_i(x, t) = \alpha_i \| f_i(x, t) \| ,$$

where the norm $\| \cdot \|$ measures the strength of the activation of f_i, and $\alpha_i > 0$ is a customizable scaling factor. Notice that the $\alpha's$ values can properly be chosen to express the interest in a specific visual feature, thus providing task-driven trajectories. We consider the gravitation field E, in which the attraction toward the distributional mass μ is inversely proportional to the squared distance from the focus of attention $a(t)$, given by

$$E(a(t), t) = -\frac{1}{2\pi} \int_{\mathbb{R}} dx \frac{a(t) - x}{\|a(t) - x\|^2} \mu(x, t) := -(e * \mu)(a(t), t), \tag{1}$$

where $*$ is the convolution operator and $e(z) = (2\pi)^{-1}(z)\|z\|^{-2}$. Once we are given the gravitational field, we can compute the Newtonian differential equation, that are

$$\ddot{a}(t) + \lambda \dot{a}(t) + (e * \mu)(a(t), t) = 0, \tag{2}$$

where dumping term $\lambda \dot{a}(t)$, with $\lambda > 0$, prevents from oscillations typical of gravitational systems and it helps to produce precise ballistic movements toward the salient target. Integrating Eq. (2) allows us to compute the visual attention trajectory at each time instant (We converted the equation to a first-order system of differential equations, as commonly done, introducing auxiliary variables. Then we used the `odeint` function of the Python SciPy library, in the setting in which it automatically determines where the problem is stiff and it chooses the appropriate integration method).

The choice of the visual features that induce the corresponding masses is determinant in modeling the behaviour of the attention system. A key property of the proposed model is that there are no restrictions on the categories of features one could exploit. While basic low-level features are considered in this work, other features associated to semantic categories (faces, objects, actions, etc.) could be introduced that might be relevant in specific visual exploration tasks. In particular, the features we consider in this paper are described as follows.

- Let $i : \mathbb{D} \to \mathbb{R}$ be the intensity of the frame, that yields the feature associated to *spatial gradient of the brightness*, $f_1 = \nabla_x i$. This features carries information about edges and, generally speaking, it reveals the presence of details in the input data.
- Let $c_j : \mathbb{D} \to \mathbb{R}$ be the color channels of the frame, with $j \in \{1, 2, 3\}$ that yields the feature associated to *spatial gradient of the color*, $f_{1+j} = \nabla_x c_j$. This features carries information about edges on the color channels and, similarly to the case of f_1, it reveals the presence of details in the input data.
- Let $o_k : \mathbb{D} \to \mathbb{R}$ be the *orientations*, that reveal the presence of edges oriented at 0°, 45°, 90° and 135°, with $k \in \{1, 2, 3, 4\}$. The feature $f_{4+k} = \nabla_x o_k$ characterizes areas oriented in a certain direction.

Please notice that here we use a natural assumption of describing the input stream with contrastive features given by the gradient function. It aims at reproducing the activity of photoreceptive cells working in color-opponent. In humans, after a reflexive shift of attention towards the source of stimulation, there is an inhibition to remain in the same location. This mechanism is called Inhibition Of Return (IOR). Originally discovered in human studies of attention, inhibition of return is a tendency for the organism to orient away from a previously attended location and biologically depends on neural structures that participate in oculomotor control[29], parietal and frontal cortex[37]. This phenomenon was first described by Posner and Cohen[38] who showed that reaction times to detect objects appearing in previously cued locations were longer than to uncued locations. The phenomenon has been demonstrated in a number of different paradigms[39,40] as part of the visual foraging behavior and may reflect a *novelty bias*[41] making visual exploration to proceed efficiently. Other authors have suggested alternative explanations of IOR depending upon task specificity such as the observation of an inhibitory effects for non-spatial attributes of irrelevant pre-cues[42,43], or computational models of negative priming[44,45] that can also generate IOR from irrelevant cues. While other implementations are equivalent and compatible with the present proposal, here we have made the choice to design the IOR in its original description, adapting it to the gravitational framework. We define a similar mechanism in our model, to prevent the trajectory to get trapped into regions of equilibrium and favour complete exploration of the scene. The dynamic of a function of inhibition $I(x, t)$ can be modeled as

$$\frac{\partial I(x,t)}{\partial t} + \beta I(x,t) = \beta g(x - a(t)), \tag{3}$$

where $g(u) = e^{-\frac{u^2}{2\sigma^2}}$ and $0 < \beta < 1$. This is directly applied to the feature masses, in order to decrease the gravitational contribution from already-visited spatial locations. As a results, the distribution of masses μ becomes

$$\mu(x,t) = \sum_i (\mu_i(x,t) - \mu_i(x,t)I(x,t))). \tag{4}$$

Scanpath prediction. Shifts on visual attention allow mammals to relocate the fovea to the next location of interest. A sequence of shifts determines a visual scanpath. The gravitational model described above provides a computational method to produce sequences of fixations and saccades, given a visual stimulus. Equation (2) provides a differential law describing attentional shifts. It can be numerically integrated to produce simulations that can be compared with data from human subjects collected by means of eye-trackers. It is worth mentioning that our model generates a continuous scanpath. The same fixation detection algorithms that are used on human recordings have been applied here to extract fixations from the output of the gravitational model. Instead, the WTA is an algorithm that provides a discrete output.

The comparison between the two models is based on similarity metrics or distances between trajectories. These metrics quantify how well the simulated sequence fit the locations visited by the human subject, taking into account also the order in which these locations are visited. In particular, results are given in terms of the following metrics:

- *String-edit distance (SED)*[33,34]. The input stimulus is divided into $m \times m$ regions, labeled with characters. Scanpaths are turned into strings by associating each fixation with the corresponding region. Finally, the string-edit algorithm[31] is used to provide a measure of the distance between the two generated strings.
- *Time-delay embeddings (TDE)*[35]. This measure is commonly used in order to quantitatively compare stochastic and dynamic scanpaths of varied lengths. It is defined as the average of the minimum Euclidean distances of each sub-sequence of length m from the original trajectory with all the possible subsequences of length m from the generated trajectory.
- *Scaled time-delay embeddings (STDE)*[34,36]. This scaled version of the previous metric is obtained by normalizing coordinates between 0 and 1, according to the size in pixels of each of the presented stimuli.

The results in Tables 1 and 2 summarized the scores (SED, TDE and STDE metrics) calculated with respect to the human scanpaths. Both models under examination, i.e. GRAV and WTA, assume that pre-attentive spatial maps are given in the system. In the comparison we include two possibilities to ensure a fair comparison. On the one hand, we follow the original implementation of the WTA and assume that a saliency map is pre-calculated and fed as input to the systems. We use the original implementation described in Itti[7]. In the second case, we use a set of more basic features, corresponding to intensity, color and orientation.

GRAV model outclasses the WTA in all cases. The results show that the introduction of the saliency map does not bring significant benefits to either model. We hypothesize that the advantage of GRAV over WTA depends on the fact that the gravitational approach allows to generate more naturalistic fixations, more centered on the center of mass of salient objects rather than drastically on the edges. This is also shown qualitatively in the Fig. 1.

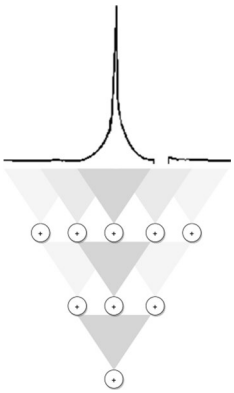

Figure 1. Example of simulated scanpath. This example shows a borderline case where the scanpath generated with WTA is unnatural because it focuses exclusively on borders with high center-surround differences. The GRAV approach, in contrast, allows to generate fixations on center of mass. Consequently, the large amount of variation in random noise on the right makes it more interesting than the square on the left.

Model	Pre-attentive maps	SED	TDE	STDE
GRAV	Basic	7.68 (0.65)	**226.70 (76.96)**	**0.80 (0.06)**
GRAV	Itti	**7.67 (0.63)**	228.08 (76.97)	0.79 (0.06)
WTA	Basic	8.41 (0.50)	425.27 (66.87)	0.65 (0.04)
WTA	Itti	8.41 (0.49)	417.12 (65.99)	0.66 (0.04)

Table 1. Results on MIT1003. In bold, the best results on average. The standard deviation values are given in round brackets. Note that the two models have performance equivalent to varying basic features. The gravitational model performs better on every metric, compared to the winnner-take-all model.

Model	Pre-attentive maps	SED	TDE	STDE
GRAV	Basic	13.81 (2.01)	**454.52 (111.17)**	**0.78 (0.04)**
GRAV	Itti	**13.77 (2.01)**	458.76 (110.79)	**0.78 (0.04)**
WTA	Basic	14.48 (2.07)	762.99 (100.94)	0.66 (0.03)
WTA	Itti	14.48 (2.07)	766.06) (101.92)	0.66 (0.03)

Table 2. Results on CAT2000. In bold, the best results on average. The standard deviation values are given in round brackets

Discussion

In the literature, with the influence of Treisman and Gelade's feature integration theory[46] and after the seminal work by Koch and Ullman[24] and Itti et al.[7], attention models are often associated with the estimation of saliency maps. Benchmarks of saliency prediction are well established[47]. Less studied is the problem of generating fixation sequences, along with the problem of explaining whether these fixations actually depend on a previous computation to combine basic features into a saliency map. In fact, such a computation seems non-trivial, albeit plausible biologically[24]. Computational models of saliency often tacitly assume that fixations can be generated with the winner-take-all mechanism[24] along with some unspecified rules of preference. Other methods for the prediction of visual attention shift have been proposed in the literature. They all assume that shift of visual attention are based on features extracted in a pre-attentive phase. In[48] the authors incorporate in the model a series of biological biases that allow for more plausible saccades, on the top of a saliency map. Even though this method delivers more precise saliency estimate, it fails to provide an explanation of the phenomenon, i.e. how these biological biases actually emerge. Other authors have developed different theories, independent of saliency (but still assuming the existence of spatial maps of features extracted in parallel in a pre-attentive phase) but their descriptions are only qualitative. This often does not allow a description of the computation underlying the visual system and, consequently, prevents a quantitative comparison. For example, Renniger et al.[49] provides an explanation of how humans explore specific artificial shapes, but it is not clear how this can be extended to a general theory of attention. Recently, data driven machine learning approaches tried to predict sequences of human fixations directly from data[50,51]. These approaches have two main flaws. The first is that their performance depends heavily on the data chosen. The second is that they fail to give a computational description of biology.

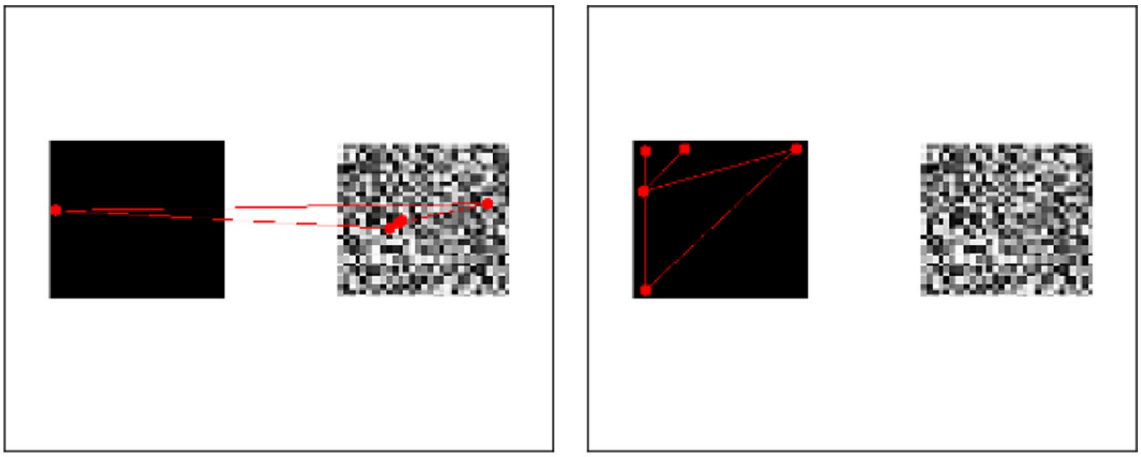

Figure 2. Field network. This network realizes the computation of a quantity proportional to the functional associate with the gravitational field. The black line on the top is illustrative. It shows a qualitative example of the distribution of cones in the retina. The maximum point correspond to the center of the fovea. A characteristic blind spot is also illustrated.

While from an application point of view they can be very useful, they say little about how human perception works. It is also unclear how these models can extend to the natural case of dynamic scenes.

The proposed gravitational approach GRAV allows to explain attentional shifts better than the current reference model, i.e. WTA. With the same features, the proposed model outperforms the winner-take-all algorithm in the task of scanpath prediction. The result is independent of the choice of the starting features, which we assume to be calculated in parallel in a pre-attentive phase. The results are more evident in the case of TDE and STDE measures, which are based on Euclidean metrics. These are, in fact, more spatially sensitive and could capture any recurring dynamics[52], such as the preference for shorter saccades. As the Fig. 1 shows qualitatively, both the distribution of the amplitudes of the saccades but also the presence of fixations in the center of objects. This claim has been quantitatively demonstrated with metrics to measure the proximity of scanpath measured with human ones. Unlike WTA, GRAV models do not strictly relay on a pre-calculated saliency map. It acts directly on feature maps that are treated as mass distributions. This ensures the versatility of the model and could directly explain how other priors, i.e. top-down priors, can intertwine in the attention mechanism, as long as they preserve the spatial conformation. Clearly, it is not the case that the eyes are purely driven by low level feature changes. Even in the absence of an explicit task, other factors such as *meaningfulness* of a location in a scene[53] can predict fixations. Similarly, it has been shown that people exhibit an understanding of *scene grammar* and move their eyes correspondingly[54], thanks to the interpretation of the scene. A preliminary attempt to interpret these high-level visual skills has been already reported in[16], where it is shown how the hidden neurons of a deep neural network trained for object classification can be integrated with a variational model, provided that the spatial map distribution is maintained . Furthermore, the GRAV model describes a continuous dynamics of the process. The output of the model presents the same step behavior of saccades that is determined by the joint contribution of the inhibition mechanism. This partially explains the advantage of the GRAV model over the WTA in terms of performance in the scanpath prediction task: when performing a saccade, in fact, the gravitational contribution of the peripheral vision continuously influences the relative positioning with respect to the final target[55]. Describing a computational model of visual attention while taking into account how this process could be implemented by a biological hardware requires considering that visual attention solves the sensory-functional trade-off between minimizing resources and organizing them efficiently to collect information, in a hierarchical structure. From this point of view, the proposed differential model has a natural interpretation in terms of local computation made by a hierarchical layers structure in which each unit is identified by a layer index $l \in L$ and a positional index $i \in \{1, \ldots, n_l\}$, where n_l is the number of units belonging to the lth layer. The first layer, $l_0 = 0$, is the system input and its units can be identified with the photoreceptors distributed on the retina. Then, $\forall l \in L - \{l_0\}$, units are defined by $u_i^l = \sum_{j \in N_i^l} \sigma(u_j^{l-1})$, where N_i^l represents the receptive field of the unit u_i^l and σ is an activation function which eventually introduce non-linearity in the computational graph. We identify two feed-forward steps for the calculation of the feedback signal encoding the eye motion command (which, in our description, is a continuous signal). This steps are, respectively, the calculation of the quantity in Eq. (1) (associated with the gravitational field) and the updating rule of the variable $\ddot{a}(t)$ (encoding the eye shifts) which derives from the differential Eq. (2). The first step is realized by a hierarchical structure (see Fig. 2) in which a layer of computational units perform a linear summation with equal weights to achieve an isotropic response. In other words, the activation function σ is linear: units at a layer l receive the activations in a neighborhood in the $l-1$ layer and propagate to the $l+1$ layer an amount of activation which is proportional to the linear summation. We will call it *field network*. Note that calculating a linear summation response with saturating non-linearities on the

inputs requires *ad hoc* adjustment of the connection weights, depending on the activity intensity and number afferent cells[56]. It is well known that the visual system is organized retinotopically and hierarchically from the retina to the visual cortex V1. The central part of the retina, the fovea, is an area of 2.5° with the best visual resolution (capacity to recognize fine details); outside this area, spatial resolution decreases sharply. In fact, while receptive field (RF) of neurons corresponding to the fovea are small, their density is high and they are overrepresented in V1, the size of RF increase and their number decrease with eccentricity. Furthermore, units of the periphery project diffusely to many central neurons which receive information from wide receptive fields[57]. The effectiveness of this neuronal organization of resources furthermore results on an enhanced visual system's effective spatial resolution[58]. The representation of stimuli with a peak function naturally implements the weighting term $\frac{a(t)-x}{\|a(t)-x\|^2} \propto \frac{1}{\|a(t)-x\|}$ within the functional action, which in that equation had a gravitational interpretation. In fact, the units connected to the receptors closer to the central retina will receive with a probability related to their distribution a greater amount of activation and, consequently, will propagate a stronger signal.

We have discussed the gravitational computations at the levels of description at which the nature of the computation itself is expressed (i.e., mathematical analysis) and at which the algorithms that implement the hierarchical computation are characterized. Now we provide a sketch of the neuronal hardware[59] that may be implementing this scheme for visual attention shifts based on attractor's laws. It is well known that simple properties are extracted from the retina to the early representation, corresponding to basic features. This representation are most likely localized within and beyond striate cortex, like the V4 for color and geometric shapes, or middle temporal and middle superior temporal areas for motion. This spatial maps are fed into the hierarchical structures and provide the input for the GRAV network. As it was proposed in the original paper of the WTA [24], a supported hypothesis is that this computations may take place in the early visual system, for example in the lateral geniculate nucleus LGN. There is in fact evidence that visual signals can travel from the periphery to the cortex and back to the LGN[60]. This hypothesys would be more in line with recent findings confirming that attention modulates visual signals before they even reach cortex by increasing responses of both magnocellular and parvocellular neurons in the LGN[61].

It is worth underlying at this step that our algorithmic formulation offers a simplified explanation of the same phenomenon, compared to the WTA circuitry described by Koch and Ullman and their proposal for a biological implementation[24]. GRAV requires elementary units to perform (weighted) linear summation, while Koch and Ullman assume (1) units to perform a max operation and (2) hypothesize the presence of a parallel network, identical in structure but performing backward calculations. The latter is introduced into their framework as a trick to retrieve the spatial location of the maximum. It is not specified, then, how this is encoded and transmitted to subsequent layers for further calculations and to generate a command signal for relocation of the fovea. In our gravitational framework, no backward computation are required. In fact, the quantity calculated by the field network will be used *as it is* to update a feedback signal that codify the fovea shift. The whole stack of computation is feed-forward and include only local computation, which makes the overall process more efficient. The neural implementation of the second step of computation in Eq. (2) requires the definition of a neuronal graph implementing an integration of the given differential Eq. (2). It is well known that a neural network implementation of differential equation is possible[62–64] and efficient in terms of execution time, compared with classical method that do not exploit parallel computing[65]. In particular, a simple implementation of finite differences methods is presented by Lee and Kung[62] together with an explicit method for calculating a general continuous and discrete neural algorithms for solving a wide range of complex partial differential equations. This scheme assumes basic functional operation that are plausibly implementable by a biological neural circuit. However, the implemetation of this second step become straightforward assuming that the dissipation term introduced in the theoretical model could be solely reduced to phenomena of friction. This derives from the purely mechanical fact of the eye residing in a plant of muscles that keep it in its natural position of *looking straight*. The dissipation term is fundamental in the proposed model to ensure precise moviments toward a target and finds its biological counterpart in the resistance to movement that derives from the plant in which the eye is placed. In this case, the updating equation would be dramatically simplified to $\ddot{a}(t) \propto -(e * \mu)(a(t), t)$, obtaining that the output of the field network on the first step directly codify a speed command to be sent to the eye muscles. Also in this case, and differently by the Koch and Ullman's proposal[24], no backward signal is necessary which gives rise a more natural implementation and efficient computing.

Finally, we notice that the dynamic nature of the model is particularly suitable for virtual reality applications. Without any modification, the proposed model can be used to navigate 360° environments. This could open the doors to a new research direction, where we emphasise the reproduction of conditions that are increasingly similar to human vision.

Methods

Datasets. Collecting eye-tracking data is a time-consuming process. Selected subjects must be invited to participate in an experiment that normally takes place at the same room, with controlled light conditions to limit the variability of the experiment, and with the need of calibrating the eye-tracking tools. In recent years, large collections of data have been made publicly available. Due to the inherently complex nature of both the stimuli and the human cognitive process, bigger eye-tracking data are necessary for a meaningful evaluation. For this reason, through all experimental evaluations of this paper, we use 2 different publicly available eye-tracking datasets. The exposure time of subjects to visual input ranges from 3 to 5 s. The number of subjects per stimulus varies from 15 to 24. Image resolution varies widely within the dataset. The details for each of the two datasets used are specified in Table 3. All images and video frames are resized to a resolution of 224×224. This makes the experiments more easily manageable, reducing the computational time. We noticed that higher resolutions did not significantly improve the performance of any of the models.

Dataset name	Details
MIT1003[9]	This dataset contains 1003 natural indoor and outdoor scenes. They are sampled with variable sizes, where each dimension is in [405, 1024]px. The database contains 779 landscape images and 228 portrait images. Fixations of 15 human subjects are provided for 3 s of free-viewing observation
CAT2000[36]	A collection of 2000 images is provided as the training portion of this dataset. Semantic content largely vary among twenty different categories. The resolution of the images is 1920 × 1080 px. Fixations of 24 human subjects are provided for 5 s of free-viewing observation

Table 3. Collection of datasets. To ensure a proper evaluation of the proposed model, a large collection of static images from two different datasets have been used. Eye-tracking data is collected in a free-viewing setup. Details are described in the right column for each of the datasets.

Feature map extraction and software implementation. The proposed gravitational model is compared with the WTA algorithm in the task of generating fixation sequences (equiv. scanpath) on a large collection of static images. Although the gravitational model has the advantage of being naturally extended to dynamic scenes (i.e., videos), we restrict ourselves to static images to make the comparison easier to evaluate. The effectiveness the gravitational model, named GRAV, in predicting saliency and scanpath on dynamic scenes has been demonstrated in a previous work[34]. We use the same input features proposed in the computational implementation of the WTA model realized by Itti[7]. Such features include an intensity channel, three color channels and four orientation channels. Notice that all the feature maps are equally weighted and no special tuning is applied, in any case, that could have improved the performance of the gravitational model. Since in its original version[24] WTA does not directly work on such basic features but on a saliency map obtained with subsequent calculation steps, both models are also evaluated while operating on the saliency map, instead of the basic features. The saliency map is generated using the code in the original implementation provided by the authors, and the implementation of the WTA model follows the description in the original paper[24]. The first fixation is chosen as the location with maximum saliency value. In the case of basic feature maps, they are first combined linearly with equal weight, then the location with maximum value on the resulting map is selected. Moreover, the selected location is inhibited within 2° of visual angle to switch attention to a subsequent location. In the original paper[24], the authors propose two additional rules for selecting subsequent locations based on proximity and similarity preference. However, they do not provide quantitative descriptions of how this should be implemented either mathematically or by biological hardware. For this reason, these rules have not been implemented in our software. It is worth pointing out that the concept of proximity preference is, instead, automatically encoded in the GRAV model that we propose in this paper. It derives from its gravitational description, where attraction is inversely proportional to the distance. More details about biological reason of such a resulting behaviour will be given in the following paragraphs.

Tuning of the parameters. The behaviour of the proposed differential model GRAV depends on a small set of parameters $\{\beta, \lambda\}$ that must be carefully selected. The parameter β was set to 0.1. We found that the choice of different values for β did not produce significantly different results. The parameter $\lambda > 0$ prevents oscillations (or orbits) because it introduces a dumping term (see, for example, the classic equation for the damped harmonic oscillator). For all the experiments we used the set of parameters that maximized the performance of GRAV in 20 images that were kept out from the described datasets for validation purposes. In particular, we performed a grid search procedure, selecting the parameters that maximized the NSS saliency score.

Accession codes

All codes used for the experiments reported in this manuscript will be made available in a public repository.

Received: 23 February 2020; Accepted: 11 September 2020
Published online: 01 October 2020

References

1. Koch, K. *et al.* How much the eye tells the brain. *Curr. Biol.* **16**, 1428–1434 (2006).
2. Borji, A., Sihite, D. N. & Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Process.* **22**, 55–69 (2013).
3. Smith, P. L. & Ratcliff, R. An integrated theory of attention and decision making in visual signal detection. *Psychol. Rev.* **116**, 283 (2009).
4. Hood, B. M., Willen, J. D. & Driver, J. Adult's eyes trigger shifts of visual attention in human infants. *Psychol. Sci.* **9**, 131–134 (1998).
5. Duncan, J. Converging levels of analysis in the cognitive neuroscience of visual attention. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **353**, 1307–1317 (1998).
6. Martinez-Conde, S., Otero-Millan, J. & Macknik, S. L. The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nat. Rev. Neurosci.* **14**, 83 (2013).
7. 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).
8. Bruce, N. & Tsotsos, J. Attention based on information maximization. *J. Vis.* **7**, 950–950 (2007).
9. Judd, T., Ehinger, K., Durand, F. & Torralba, A. Learning to predict where humans look. In *IEEE 12th International Conference On Computer Vision* 2106–2113 (2009).
10. Zanca, D. & Gori, M. Variational laws of visual attention for dynamic scenes. In *Advances in Neural Information Processing Systems* 3823–3832 (2017).

11. Cornia, M., Baraldi, L., Serra, G. & Cucchiara, R. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)* 3488–3493 (IEEE, 2016).
12. Borji, A. & Itti, L. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 185–207 (2013).
13. McMains, S. A. & Kastner, S. *Visual Attention* 4296–4302 (Springer, Berlin, 2009).
14. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194 (2001).
15. Connor, C. E., Egeth, H. E. & Yantis, S. Visual attention: bottom-up versus top-down. *Curr. Biol.* **14**, R850–R852 (2004).
16. Zanca, D., Gori, M. & Rufa, A. A unified computational framework for visual attention dynamics. *Prog. Brain Res.*https://doi.org/10.1016/bs.pbr.2019.01.001 *(2019).*
17. Hankinson, G. The brand images of tourism destinations: a study of the saliency of organic images. *J. Product Brand Manag.* **13**, 6–14 (2004).
18. Milosavljevic, M., Navalpakkam, V., Koch, C. & Rangel, A. Relative visual saliency differences induce sizable bias in consumer choice. *J. Consum. Psychol.* **22**, 67–74 (2012).
19. Guo, C. & Zhang, L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.* **19**, 185–198 (2009).
20. Sitzmann, V. *et al.* Saliency in VR: how do people explore virtual environments?. *IEEE Trans. Vis. Comput. Graph.* **24**, 1633–1642 (2018).
21. Womelsdorf, T., Anton-Erxleben, K., Pieper, F. & Treue, S. Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat. Neurosci.* **9**, 1156 (2006).
22. Corbetta, M. *et al.* A common network of functional areas for attention and eye movements. *Neuron* **21**, 761–773 (1998).
23. Nobre, A. C. *et al.* Functional localization of the system for visuospatial attention using positron emission tomography. *Brain J. Neurol.* **120**, 515–533 (1997).
24. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence* (ed. Vaina, L. M.) 115–141 (Springer, Dordrecht, 1987).
25. Duan, H. & Wang, X. Visual attention model based on statistical properties of neuron responses. *Sci. Rep.* **5**, 8873 (2015).
26. Itti, L. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vis. Cogn.* **12**, 1093–1123 (2005).
27. Zhang, X., Zhaoping, L., Zhou, T. & Fang, F. Neural activities in v1 create a bottom-up saliency map. *Neuron* **73**, 183–192 (2012).
28. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607 (1996).
29. Westerberg, J. A., Maier, A. & Schall, J. D. Priming of attentional selection in macaque visual cortex: feature-based facilitation and location-based inhibition of return. *Eneuro* **7**, 1–15 (2020).
30. Burkhalter, A. & Bernardo, K. L. Organization of corticocortical connections in human visual cortex. *Proc. Natl. Acad. Sci.* **86**, 1071–1075 (1989).
31. Jurafsky, D. & Martin, J. H. *Speech and Language Processing* Vol. 3 (Pearson, London, 2014).
32. Brandt, S. A. & Stark, L. W. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *J. Cogn. Neurosci.* **9**, 27–38 (1997).
33. Foulsham, T. & Underwood, G. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *J. vis.* **8**, 6–6 (2008).
34. Zanca, D., Melacci, S. & Gori, M. Gravitational laws of focus of attention. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
35. Wang, W. *et al.* Simulating human saccadic scanpaths on natural images. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 441–448 (IEEE, 2011).
36. Zanca, D., Serchi, V., Piu, P., Rosini, F. & Rufa, A. Fixatons: a collection of human fixations datasets and metrics for scanpath similarity. ArXiv preprint, arXiv:1802.02534 (2018).
37. Bichot, N. P. & Schall, J. D. Priming in macaque frontal cortex during popout visual search: feature-based facilitation and location-based inhibition of return. *J. Neurosci.* **22**, 4675–4685 (2002).
38. Posner, M. I., Rafal, R. D., Choate, L. S. & Vaughan, J. Inhibition of return: neural basis and function. *Cogn. Neuropsychol.* **2**, 211–228 (1985).
39. Gibson, B. S. & Egeth, H. Inhibition and disinhibition of return: evidence from temporal order judgments. *Percept. Psychophys.* **56**, 669–680 (1994).
40. Pratt, J. & Abrams, R. A. Inhibition of return in discrimination tasks. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 229 (1999).
41. Milliken, B. & Tipper, S. P. Attention and inhibition. In *Attention* (ed. H. Pashler) 191–221 (Psychology Press, 1998).
42. Mondor, T. A., Breau, L. M. & Milliken, B. Inhibitory processes in auditory selective attention: evidence of location-based and frequency-based inhibition of return. *Percept. Psychophys.* **60**, 296–302 (1998).
43. Law, M. B., Pratt, J. & Abrams, R. A. Color-based inhibition of return. *Percept. Psychophys.* **57**, 402–408 (1995).
44. Houghton, G. & Tipper, S. P. *A Model of Inhibitory Mechanisms in Selective Attention* (Academic Press Ltd, London, 1984).
45. Milliken, B., Tipper, S. P., Houghton, G. & Lupiáñez, J. Attending, ignoring, and repetition: on the relation between negative priming and inhibition of return. *Percept. Psychophys.* **62**, 1280–1296 (2000).
46. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
47. Bylinskii, Z. *et al.* Mit saliency benchmark. (Accessed 1 September 2019); http://saliency.mit.edu/.
48. Le Meur, O. & Coutrot, A. Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vis. Res.* **121**, 72–84 (2016).
49. Renninger, L. W., Coughlan, J. M., Verghese, P. & Malik, J. An information maximization model of eye movements. In *Advances in Neural Information Processing Systems* 1121–1128 (2005).
50. Jiang, M. *et al.* Learning to predict sequences of human visual fixations. *IEEE Trans. Neural Netw. Learn. Syst.* **27**, 1241–1252 (2016).
51. Kümmerer, M., Wallis, T. & Bethge, M. Deepgaze ii: Predicting fixations from deep features over time and tasks. In *17th Annual Meeting of the Vision Sciences Society (VSS 2017)* 1147–1147 (2017).
52. Abarbanel, H. D., Carroll, T., Pecora, L., Sidorowich, J. & Tsimring, L. Predicting physical variables in time-delay embedding. *Phys. Rev. E* **49**, 1840 (1994).
53. Henderson, J. M. & Hayes, T. R. Meaning guides attention in real-world scene images: evidence from eye movements and meaning maps. *J. Vis.* **18**, 10. https://doi.org/10.1167/18.6.10 (2018).
54. Vo, M.L.-H., Boettcher, S. E. & Draschkow, D. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Curr. Opin. Psychol.* **29**, 205–210 (2019).
55. Veneri, G., Federighi, P., Rosini, F., Federico, A. & Rufa, A. Spike removal through multiscale wavelet and entropy analysis of ocular motor noise: a case study in patients with cerebellar disease. *J. Neurosci. Methods* **196**, 318–326 (2011).
56. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019 (1999).
57. Carpenter, R. Movement control: moving the mental maps. *Curr. Biol.* **5**, 1082–1084 (1995).
58. Anton-Erxleben, K. & Carrasco, M. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat. Rev. Neurosci.* **14**, 188 (2013).
59. Marr, D. & Poggio, T. *From Understanding Computation to Understanding Neural Circuitry* (MIT Press, Cambridge, 1976).

60. Briggs, F. & Usrey, W. M. A fast, reciprocal pathway between the lateral geniculate nucleus and visual cortex in the macaque monkey. *J. Neurosci.* **27**, 5431–5436 (2007).
61. McAlonan, K., Cavanaugh, J. & Wurtz, R. H. Guarding the gateway to cortex with attention in visual thalamus. *Nature* **456**, 391–394 (2008).
62. Lee, H. & Kang, I. S. Neural algorithm for solving differential equations. *J. Comput. Phys.* **91**, 110–131 (1990).
63. Lagaris, I. E., Likas, A. & Fotiadis, D. I. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Netw.* **9**, 987–1000 (1998).
64. Tsoulos, I. G., Gavrilis, D. & Glavas, E. Solving differential equations with constructed neural networks. *Neurocomputing* **72**, 2385–2391 (2009).
65. Yadav, N., Yadav, A. & Kumar, M. *Neural Network Methods for Solving Differential Equations* 43–100 (Springer, Dordrecht, 2015).

Acknowledgements

We thank Frédéric Precioso and Lucile Sassatelli for fruitful discussions on the model which stimulated a wider view on the topic as well as interesting application perspectives in computer vision.

Author contributions

D.Z., M.G. and S.M. contributed to the mathematical formulation of the model. D.Z. and A.R. contributed to the analysis of the biological plausibility. D.Z. conceived and conducted all the experiments. All authors analysed the results and reviewed the manuscript.

Funding

The work was partially supported by RoNeuro and Liquidweb srl who participate in the Neurosense Joint Lab of the Dept. Medicine Surgery and Neuroscience, University of Siena.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.